Journal of Economic Literature
Vol. XXXIV (September 1996), pp. 1236–1263

Reimbursing Health Plans and Health Providers: Efficiency in Production Versus Selection

JOSEPH P. NEWHOUSE
Harvard University

I would like to thank Ernie Berndt, David Cutler, Peter Diamond, Randall Ellis, William Encinosa, Richard Frank, Jerry Green, Albert Ma, Thomas McGuire, Mark Pauly, Kathy Swartz, Richard Zeckhauser, and anonymous referees for helpful comments on a preliminary draft. Naturally I am responsible for the remaining infelicities.

I. *Introduction*

ONE OF THE BEST KNOWN tradeoffs in insurance economics is between risk aversion and moral hazard. Greater insurance coverage implies less risk bearing by the insured but induces greater moral hazard. As a corollary, the less demand responds to price, the greater should be the coverage of the loss (Zeckhauser 1970; Issac Ehrlich and Gary Becker 1972; Pauly 1986).

The tradeoff between risk aversion and moral hazard has given rise to much health economics literature on the desirability of cost sharing in health insurance, i.e., the price to the insured at the time of use, and we now know a good bit empirically about this tradeoff (Newhouse and the Insurance Experiment Group 1993, ch. 4). Perhaps based on these findings, many now believe that some initial cost sharing is optimal; the outcome should not be the corner solution of no cost sharing (e.g., U.S. Bipartisan Commission on Comprehensive Health Care 1990, p. 63).

This paper argues that widespread health insurance creates another important tradeoff, less well recognized in the literature, between efficiency in production and selection. By efficiency in production I mean least cost treatment of a patient's medical problem, holding quality constant. Thus, efficiency includes the quantity of services used to treat the problem, as well as the unit price of those services. By selection, I mean actions of economic agents on either side of the market to exploit unpriced risk heterogeneity and break pooling arrangements, with the result that some consumers may not obtain the insurance they desire.

[1] Note the analogy with Ramsey pricing. Full coverage of a potential loss is optimal under the following conditions: no moral hazard; actuarially fair insurance; and a risk averse consumer. Full insurance above a deductible is optimal with positive loading charges but no moral hazard (Kenneth Arrow 1963). See also the exchange between Arrow (1968) and Pauly (1968).

[2] Any reasonable amount of risk aversion and administrative cost will rule out the optimality of the other corner solution, that of no insurance; Newhouse and the Insurance Experiment Group (1993, ch. 4) estimate the welfare loss from no coverage.

Whereas the essence of the moral hazard-risk aversion tradeoff is captured by the cost that the patient bears at the time of use, e.g., the size of a deductible, the essence of the selection-efficiency tradeoff is captured by the cost the health plan or medical provider bears at the time of use, or the amount of supply-side cost sharing, to use the term of Ellis and McGuire (1993). Analogous to coinsurance on the demand side, supply-side cost sharing in its simplest form is a linear combination of fee-for-service and capitation pricing, but nonlinear schedules are obviously possible. This paper largely concerns the tradeoff between these two bases of pricing in both the health insurance and the medical care delivery markets.

The paper is organized as follows. Section II describes changes in the structures of the health insurance and medical care industries and how those changes have affected the tradeoff between efficiency in production and selection. Section III delineates a theory of selection based on the traditional Rothschild-Stiglitz model, but shows that modifying an assumption reverses the conclusions of that model in a way that better accords with what is observed, namely limited pooling in a competitive insurance market. The next several sections take up efficiency in production. They show how modifying assumptions of the yardstick competition model changes its conclusion that fully prospective pricing is optimal to a conclusion that a mixed or non-fully prospective basis of pricing is desirable. Selection considerations strengthen that conclusion. The final sections consider the possibility that nonprice regulation could improve the selection-efficiency tradeoff, the empirical literatures on selection and prospective pricing of different risks, and policy implications.

II. *Medical Care and Health Insurance: Descriptive Material and a First Pass at the Selection-Efficiency Tradeoff*

The Traditional System: Selection. Until roughly the 1980s medical care delivery and health insurance were two distinct industries in the United States.[3] Medical care delivery was priced on the basis of fee-for-service; physicians received a fee for each service, and hospitals were paid for each day plus additional payments for ancillary services, such as the operating room. Some insurance plans reimbursed hospitals' "cost," which for these purposes is analytically similar to a fee (Patricia Danzon 1982).

Insurance was and still is purchased largely through employers, who subsidized it, or was provided through the large government programs of Medicare and Medicaid. Employers frequently offered their employees only one insurance plan. Medical costs were generally passed on from the "insurer" to the employer (experience rating) with little review, so that the insurer bore little risk.[4] The government was also mostly passive, paying the claims submitted. Demand was limited by demand-side cost sharing and by excluding certain services from coverage. Price competition among insurers centered on loading charges.[5]

This system minimized the incentive of both insurers and providers to select

[3] Many other countries such as Canada, France, Germany, and Japan also separate insurance and delivery of care, and they also share to some degree fee-for-service reimbursement. Some of these countries pay hospital physicians by salary, emphasizing the diversity of payment bases.

[4] With experience rating, the risk was that the employer would choose to insure through another insurer in the subsequent year, so that any shortfall in the current year could not be recouped.

[5] Loading charge is the excess of premium over payments to providers. After the mid 1970s self-insurance became the norm among large and middle-size employers, though they contracted with insurance companies for administrative services, thus formalizing the competition on loading charges.

against bad risks. Physicians and hospitals were paid more for additional services that sicker patients required; insurers in turn passed these costs on to employers, who largely paid them with little question.

The Traditional System: Production Efficiency. For services widely covered by insurance, the insurer could not observe a meaningful market price. Hence, the supply price was set in an administrative transaction between the insurer and the provider. Insurance contracts usually entitled consumers to seek care from any provider on the same terms, a clear barrier to price competition among providers. Even if there was some demand-side cost sharing, insurance weakened consumers' incentives to search.

As a result, there was every reason to believe supply prices exceeded competitive levels. Denote the competitive fee by C, a fee at the level of marginal cost associated with the competitive quantity and quality of service. Pauly (1980) showed that fees above C^* gave providers an incentive to induce imperfectly informed patients to consume more services than fully informed consumers and that the incentive to overservice increased with the fee. Consistent with this prediction, Mark Chassin et al. (1987) found that a sixth to a third of three commonly performed procedures in the fee-for-service system provided zero or negative clinical benefit.

Thus, the traditional system appeared to produce the treatment of a given medical problem inefficiently.[10] The tradeoff between selection and production efficiency tilted sharply in favor of minimizing selection.

The Evolving System: Vertical Integration. For many years so-called Health Maintenance Organizations (HMOs), which integrate the provision of insurance and medical care services, coexisted with the dominant traditional system.[11] HMOs contracted with consumers to provide "necessary" medical services in return for a capitation (a lump sum per person per month). Typically they employed or contracted with a set of physicians; they also contracted with or even owned hospitals. They thereby tried to reduce moral hazard by controlling the quantity of services (James Baumgardner 1991). If one sought care from a physician outside the HMO, one typically paid out-of-pocket. Until recently there was either no or only one HMO in a local market; perhaps for that reason the presence of HMOs did not bring much price competition to the delivery system.

In recent years such vertical integration of insurance and delivery has spread

[6] In the case of traditional Medicare, for example, the prices facing consumers for hospital care were and still are independent of the hospital chosen. This was also true for many privately insured patients who had no cost sharing at the margin because of first-dollar coverage or because they exceeded a stop-loss limit.

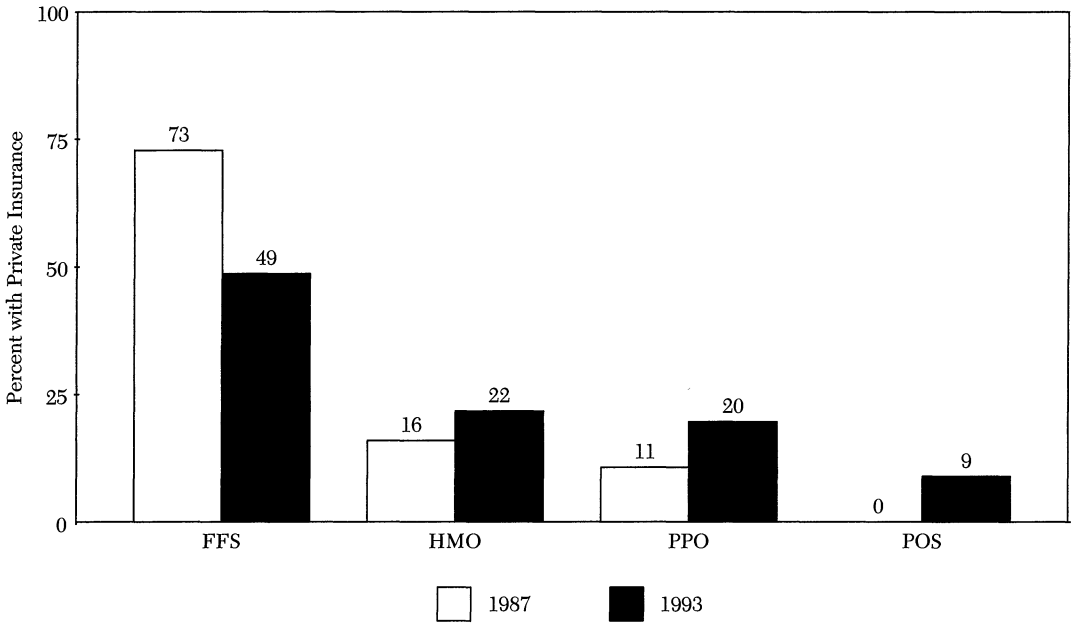
[7] Strictures on price advertising raised the costs of search.

[8] Medical ethics (primum non nocere—above all, do no harm) may reduce overservicing, but few economists believe that financial incentives are irrelevant to clinical decision making (Pauly 1980). Potential detection may also reduce overservicing through the sanction of loss of business (Michael Darby and Edi Karni 1973; David Dranove 1988); analogously it would reduce underservicing in capitation arrangements.

[9] For a review and additional evidence see Robert Brook and Elizabeth McGlynn (1991). Brook (1993) points out that these seemingly high rates of zero and negative benefit procedures are found outside the United States as well; in these countries, however, there is also little incentive to produce efficiently.

[10] Criticisms of the fee-for-service system had come much earlier and not just from economists. George Bernard Shaw wrote: "That any sane nation, having observed that you could provide for the supply of bread by giving bakers a pecuniary interest in baking for you, should go on to give a surgeon a pecuniary interest in cutting off your leg, is enough to make one despair of political humanity." (*The Doctor's Dilemma*, 1911)

[11] By far the largest example for many years was the Kaiser Health Plan.



FFS = fee-for-service; HMO = no reimbursement for services from non-HMO providers; PPO, POS = lower cost for services from network providers.

Figure 1. Growth of Managed Care, 1987-1993

Source: Health Insurance Association of America Surveys

under the rubric of "managed care" (Figure 1). Typically consumers face little or no out-of-pocket payment if they choose a provider from the integrated plan's network.[12] Plans, of course, may use both price and quality to select network providers.

The spread of managed care means that providers not in any network can find it difficult to attract patients.[13] This has set up a demand from physician groups for "any-willing-provider" legislation, meaning that any provider willing to meet the price terms of the health plan must be accepted into its network. I return to this policy issue in the concluding section.

The insurance market has also changed. Over half of American employees now have a choice of insurance plan at the place of work, with the employee paying part or all of a plan's incremental premium (Table 1). Thus, unlike the traditional health insurer for whom price competition was over loadings, the integrated plan now faces price competition on the costs of medical care. As a result, plans have stronger incentives than before to produce efficiently—as well as to select good risks.

Incentives for hospitals and physicians are also changing. Since fiscal 1984, Medicare, which accounts for about 30 percent of hospital revenue, has paid hospitals a lump sum per admission through the Prospective Payment System

[12] Unlike traditional HMOs, plans now often partly reimburse if the consumer goes outside the network. Traditional HMOs have also now started to do this with so-called point-of-service options.

[13] This implies excess supply in the physician market.

TABLE 1
NUMBER OF HEALTH PLAN CHOICES FOR PRIVATE SECTOR EMPLOYEES, 1993

Number of Plans Offered per Establishment	Weighted by Establishments	Weighted by Number of Employees
1	76%	48%
2	16%	23%
3	5%	12%
4	2%	6%
5 or more	1%	11%

Source. Unpublished data, courtesy of Stephen Long.

(PPS), strengthening hospitals' incentive to produce efficiently but also to avoid high cost patients. Some other payers also pay per admission.

In addition to facing greater price competition, network physicians now may have contracts in which they receive bonuses (pay penalties) if the services they order in a given time period fall short of (exceed) a target amount. Such contracts may result in marginal fees below the competitive level C^*; capitation of the physician is the extreme case, because marginal revenue is zero.

If fees paid by a plan are below C^*, Pauly (1980) described the conditions under which production might still be efficient: competition among plans; and sufficiently well informed patients to detect underservice. Of course, that latter assumption is in some doubt because of the evidence of overservicing in the traditional system. Thus, to the degree that patients are poorly informed, physicians with such contracts have an agency problem that is opposite from that of the traditional system; if the marginal revenue of a service is less than marginal cost, they have an incentive to underservice.[15] Moreover, to the degree physicians receive the same capitation for each patient, they have an incentive to avoid high cost patients.

All these developments suggest that both efficiency and selection may now be greater than in the traditional system.

Plan Pricing and Contract Length. Most health insurance plans, whether integrated or not, do not charge a premium that equals the expected cost of an individual or family. For example, a plan generally receives the same premium for all single employees of a given firm who choose that plan, even though some may have a chronic disease that increases their expected cost to the plan and others do not.

I therefore assume that transaction costs preclude plans' pricing at an individual's expected cost, both in private insurance markets and in public programs such as Medicare. Instead, the plan's premium is a function of the average risk in a heterogeneous group, and risk-bearing plans enroll both profitable and unprofitable members. As a result, they have an incentive to shun ("dump") unprofitable patients and to attract ("cream") profitable ones. The PPS gives a hospital the same incentive for the same reason. Moreover, plans have an incentive to structure contracts with hospitals and physicians to use underservice to

[14] The distinction between the PPS, which pays lump sums that vary by Diagnosis Related Group (DRG), and fee-for-service can be overdrawn because several DRGs are defined using a procedure, for example cardiac catheterization (Mark McClellan 1993).

[15] Tort law would appear relatively ineffectual at redressing this incentive, because few cases of medical negligence result in claims, probably under five percent (Paul Weiler et al. 1993).

[16] Similarly, to the degree an employer self-insures medical care services, the employer generally will not pass on to individual wages any individual differences in expected health care costs; my wage in general will not differ from yours because I have high blood pressure and therefore are at higher risk of hospitalization and you do not. Still Jonathan Gruber (1994) finds evidence that mandated maternity benefits differentially affected wages of women in childbearing years as a group.

motivate selective disenrollment. If all plans have this incentive, unprofitable patients may find all plans underserve them or otherwise try to induce disenrollment; in effect, such patients may be unable to obtain the desired insurance because of selection.

The length of the insurance contract is typically a year. Consequently, to the degree that poor risks pay more for insurance, as on average they do (despite the uniform pricing by plan within group), individuals cannot buy insurance against becoming a bad risk in the future, or, in a family context, against having a child who is a bad risk (Diamond 1992).

The Selection-Efficiency Tradeoff: A First Pass. A fully prospective payment for a bundle of services yields technically efficient production because the firm, whether an integrated health plan or a hospital or a physician, entirely keeps any residual; as a result, rents in factor prices do not induce excess services, and managerial effort is optimal. Uniform fully prospective payment for a heterogeneous group of persons, however, gives the firm the maximum incentive to select good risks and avoid bad ones. This implies a tradeoff between selection and production efficiency (Ma 1994).

In terms of supply prices, the tradeoff between efficiency and selection has two corner solutions: no supplier cost sharing, corresponding to the traditional fee-for-service system assuming fees at or above marginal cost, and full supplier cost sharing at the margin, for example a capitation or fully prospective payment, whether paid to the health plan, hospital, or physician.

Just as plans can be treated as agents for consumers, providers can be treated as agents for plans. In this essay I therefore analyze the selection-efficiency tradeoff at the level of plans, hospitals, and physicians similarly. By fee-for-service reimbursement in the insurance plan market, I mean a body above the competing plans that partially reimburses them for the services their enrollees use. This body could be the government or an employer or coalition of employers that redistributes a pool of premium dollars among plans; the health alliance in the Clinton health reform proposal was such a body.

As noted above, many believe that on the demand side neither no insurance nor full insurance is likely to be optimal. This essay argues that a similar conclusion applies to the supply side; neither corner solution is likely to be optimal.

III. *A Theory of Selection*

I begin from Michael Rothschild and Joseph Stiglitz' (1976) well known model, exposited recently in this journal by Nicholas Barr (1992), and then proceed to modify it.[20] The modification alters two conclusions of the model in the direction of greater realism: 1) In the

[17] On the other hand, the spot-price-like pricing system preserves an incentive not to become a bad risk (Ehrlich and Becker 1972). I treat that incentive as unimportant at the margin in the discussion of an optimal tradeoff below (Henry Aaron 1994).

[18] Rents in factor prices, however, will lead the firm to substitute away from the factor(s) with the rents.

[19] An insurer-determined physician fee schedule at or above C^* leaves the physician bearing any increase in the unit cost of the service, but no cost sharing for the number of units ordered to treat a problem.

[20] Charles Wilson (1977) proposed a similar model but obtained a pooling equilibrium by abandoning Rothschild-Stiglitz' assumption of Nash behavior (other insurers' behavior is given). Instead, he substituted the assumption that insurer i assumed, when making its offer, that other insurers would immediately withdraw any contracts made unprofitable by i's offer. I have not followed Wilson's lead because I regard Nash behavior as a reasonable approximation to the traditional insurance market. In the future, if local markets come to be dominated by a few large HMOs, Wilson's assumption may be more applicable.

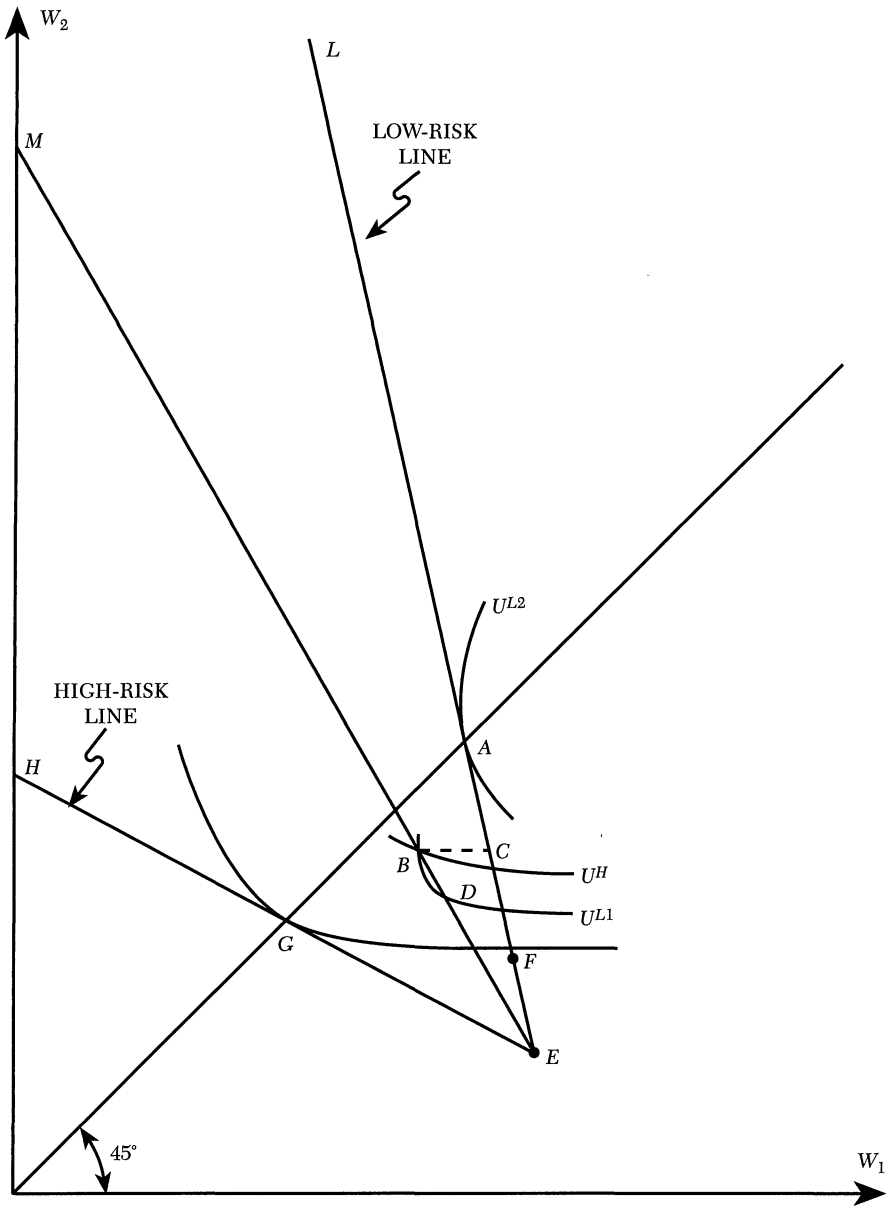


Figure 2. Separating and Pooling Equilibria

Rothschild-Stiglitz model a pooling equilibrium, i.e., heterogeneous risks in the same insurance plan, cannot exist. Yet large HMOs appear to be reasonably stable organizations (i.e., in equilibrium) and clearly enroll heterogeneous risks at the same premium. Thus, reality suggests some pooling occurs. 2) In the Rothschild-Stiglitz model high-risk persons obtain the insurance they wish, and low-risk individuals do not. In reality, however, it is high-risk, not low-risk individuals who tend to have trouble obtaining the desired amount of insurance, for

example because of pre-existing condition clauses or industry redlining.

The standard Rothschild-Stiglitz diagram is shown in Figure 2; three budget lines are shown through E, the consumer's endowment in the absence of insurance. EL is a fair-odds line for low-risk persons, the rate at which a low-risk person can trade money (pay a premium) in the no accident state (1) to receive money in the accident state (2). EL has slope $-(1-p_L)/p_L$, where p_L is the probability of an accident with costs c happening to a low-risk individual. Similarly, EH is a fair-odds line for high-risk persons with slope $-(1-p_H)/p_H$, where p_H is the probability of an accident with costs c happening to a high-risk person, $p_H > p_L$.

One can interpret the middle line EM in two ways. First, following Rothschild and Stiglitz, it could represent the fair-odds line for a population with two types of risks, reflecting the population probability of an accident $\bar{p} = (1-\lambda)p_L + \lambda p_H$, where λ is the proportion of high-risk individuals in the population. But we shall want another interpretation below: the price line facing a low-risk person if the insurance has a positive loading charge.

To show that a pooling equilibrium is not possible, consider point B and interpret EM as the population fair-odds line. The two indifference curves through B, U^H and U^{L1}, correspond to the two risk types. Rothschild-Stiglitz assume the consumer maximizes expected utility, $E(U) = (1-p)U(W_1) + pU(W_2)$, where W_1 is wealth with no accident and W_2 is wealth with an accident. Thus, the slope of risk type i's indifference curve at B is $U_{W1}/U_{W2} = (1-p_i)U'/p_iU' = (1-p_i)/p_i$, where i indexes risk type.[23] Because $p_H > p_L$, the high risks' indifference curve at B will be flatter than the low risks', thereby opening a wedge to the southeast of B.

In principle, a competitive insurer could offer a contract D lying in this wedge that would attract low risks but not high risks, because it would lie above the low-risk indifference curve U^{L1} but below the high-risk indifference curve U . At B an insurer who attracts a representative sample of risks breaks even because B is on the population fair-odds line; hence, if D is in the neighborhood of B and attracts only low risks, it will make a profit. Thus, or so goes the argument, in a competitive market some insurer will offer D and break a pooling equilibrium.

This result, however, depends upon the implicit assumption that offering D is costless. Suppose drawing up a contract that disproportionately attracts low risks is costly. If those contracting costs are large enough, such a contract might not be offered, and a pooling equilibrium might still exist.

Such a pooling equilibrium must be on the pooled budget line. Where on that line would the equilibrium be? Assume for convenience that the pooled plan has

[21] The straight line of the diagram requires that the loading proportion be constant as a function of the size of the monies to be transferred to the accident state.

[22] This result does not depend upon asymmetric information.

[23] This assumes the two risk types have the same utility function and that utility is separable in health and wealth.

[24] As is well known, if there is an equilibrium in the standard Rothschild-Stiglitz model with asymmetric information, it is a separating equilibrium yielding G for high risks (full insurance) and F for low risks (the point on their fair-odds line closest to full insurance that does not attract high risks).

[25] Joachim Neipp and Zeckhauser (1985) have also addressed the question of why one observes insurance plans with heterogeneous risks. They introduce the notion of stickiness among consumers choosing an insurance plan; once in a plan enrollees tend to remain. Plans, however, might well be able to overcome such stickiness by incurring costs; such costs would be analytically similar to contracting costs.

no loading charges. To the low-risk population the pooled fair-odds line does not look actuarily fair but rather like a policy with a proportional loading charge. With no loading, low risks want full insurance (point A). If the low-risk population has to buy on the pooled odds line, however, it wants less than full insurance (Arrow 1963). Suppose without loss of generality that on the pooled odds line the low-risk population prefers B. For the high-risk population B is a corner solution. They wish more insurance (to move up the odds line), but no one offers them such a contract because by assumption low risks do not buy, preferring B. Hence, any contract above B is bought only by high risks and loses money. Thus, B represents an equilibrium; low risks are at their preferred point, and high risks are at their most preferred feasible point.

With this result we can also show how any fixed costs of writing separate insurance contracts affect the degree of pooling. Relative to B, if only low risks were in the plan, each would have paid a premium in excess of actuarial value equal to the horizontal distance between the population-odds line and the low-risk-odds line, or $C - B$. Thus, if the cost of a separate contract is larger than $C - B$ times the number of low risks, it will not pay to move from the pooled equilibrium at B to one that segregates the low risks, and the equilibrium at B will be maintained. The size of $C - B$ is a function of the distance between EL and EH, or the difference between p_L and p_H.[29] Thus, the closer are p_L and p_H, the less the contracting costs must be to support pooling.

Note how the introduction of sufficiently large contracting costs has stood the standard Rothschild-Stiglitz conclusion on its head. If there was an equilibrium in their model, high risks were at their most preferred point and low risks were at their most preferred feasible point. Here it is just the opposite.

The importance of contracting costs in practice will depend in part on the institutional setting. If insurance plans can, for example, offer different benefits, such as covering or failing to cover mental health services or offering plans with various deductibles, transaction costs of separate contracts should be low. This realization has led some to endorse legislation requiring standardized benefits (Alain Enthoven 1988).

But even if such legislation were in place and enforceable, plans would still have many other tools to separate risks and so substantial pooling is unlikely in a

[26] Because the slope of the fair odds budget line for a given risk class is $-(1 - p_i)/p_i$, it follows that at the point of tangency with the budget line $U'(W_1) = U'(W_2)$, or that wealth in the two states is equal; thus, full insurance (equal wealth regardless of state of the world) is optimal if there are no loading charges.

[27] I follow Rothschild-Stiglitz and assume that in equilibrium all contracts make zero profit. Encinosa has pointed out to me that if this assumption is relaxed to a zero-profit condition on the ensemble of contracts, low risks may be able to subsidize high risks and make both better off, as in the Rothschild-Stiglitz model.

[28] Note that for any given amount of fixed costs, the greater the absolute number of people in the market, ceteris paribus, the more restrictive the pooling equilibrium, analogous to Adam Smith's dictum that the division of labor is limited by the extent of the market. For ease of exposition, I have followed Rothschild and Stiglitz and ignored loading costs in the separating equilibrium case. In reality all contracts will carry loading costs; the text result requires additional costs for separating contracts.

[29] For gauging the size of fixed costs the number of low risks obviously matters. That number increases with the proportion of low risks in the population, but such an increase has an offsetting effect of moving the pooled-risk-odds line closer to EL.

[30] Pauly (1986) tended to minimize the importance of asymmetric information, arguing that the Rothschild-Stiglitz prediction, that bad risks would get what they wanted but good risks would not, was the opposite of what was usually observed. The change in assumptions discussed in the text may be the explanation.

setting of individual plan choice.[31] Nonetheless, many possible tools for segregating risks imply higher contracting costs than simply writing different benefit language. For example, a plan might try to give physicians incentives to be rude to high-cost patients to induce disenrollment, but physicians could consider this a violation of medical ethics. Similarly, many physicians might balk at feigning uncertainty about how to treat a patient's problem to convince the patient to go elsewhere. In any event, perhaps because of contracting costs, one observes some pooling. But the extreme skewness of medical spending—in any one year, the top five percent of spenders account for half the spending—emphasizes the money at stake for plans that insure bad risks at average rates; I examine the evidence on selection below.[32] And the well-known increase in medical care costs has upped the reward for selection.

To get at the optimal selection-efficiency tradeoff, one must characterize the welfare loss of any inability to buy the desired insurance because of increased selection. One can approximate the value of risk aversion from insurance (Newhouse and the Insurance Experiment Group 1993, ch. 4), so the issue becomes how the likelihood of a person's not obtaining the desired insurance changes as the incentives for selection increase.

Because of fixed costs in establishing an integrated plan, only a limited number of plans will serve a local market. If the number of plans were fixed, as the rewards from selection rise, the likelihood that some individuals will not find a welcoming plan will increase. Moreover, depending on the distribution of risks, as the incentives to select rise, the marginal number of disfavored people could increase over some range. On the other hand, if the number of plans increased with stronger incentives to select, the degree of risk segmentation among them would increase, which would tend to diminish the ability to insure against becoming a future bad risk. Again the marginal welfare loss from selection could increase as the degree of prospectivity rises. I return to these points in Section IX in the context of an interior equilibrium with respect to the basis of payment.

IV. *Efficiency in Production: Yardstick Competition*

The traditional system largely avoided selection problems by minimizing price competition on medical care costs and by using fee-for-service as a payment mechanism. Why then is fee-for-service payment not universal?

A different strand of economic literature than the selection literature has focused on incentives to produce efficiently, and some of it is unsympathetic to fee-for-service and more sympathetic to fully prospective or "high-powered" payment schemes. This literature contrasts with the selection literature in two ways. Whereas the selection literature focused on the heterogeneity of consumers with respect to expected spending (i.e., high- and low-risks), this other literature focuses on the heterogeneity of firms with respect to production costs. Whereas the selection literature focused on asymmetric information between consumers and insurers, the other literature focuses on asymmetric information about

[31] In the traditional system pooling came about because a group formed for other reasons, namely employees at a firm, purchased insurance. Such pooling can break down in small firms, consistent with the lower rate of insurance there, and as well among groups without a connection to the labor market, one rationale for Medicare and Medicaid (George Akerlof 1970).

[32] It is, however, the lesser skewness in expected spending, not actual spending that is relevant.

managerial effort between the firm and a regulator.

A well-known model from this other literature, Andrei Shleifer's (1985) model of yardstick competition, can serve as a device for organizing the next few sections. In Shleifer's model, as in many models in this literature, the cost of production, which is observable to a regulator, is partly a function of unobservable managerial effort. Shleifer makes the key assumption that all firms produce a homogeneous product, and concludes that each firm will produce efficiently if paid at the average of all other firm's marginal costs. If marginal costs are less than average costs, he specifies a lump-sum transfer so the firm breaks even.

Shleifer's important point is simply that managers will invest the socially optimal amount of effort if they keep the full reward from doing so. Any price that is independent of the firm's cost has this property and gives each firm an incentive to reveal its optimal cost structure.

Shleifer uses payments to hospitals through the PPS to illustrate yardstick competition. Much of the health economics literature subsequent to Shleifer has also dealt with high- vs. low-powered payments to hospitals, but high- vs. low-powered payments to HMOs for all medical services are analytically similar. Following the literature, the ensuing discussion generally uses the context of payment for hospital admissions and speaks of a generic regulator. In the hospital context the regulator could be a public or private insurance plan or an HMO that contracts with hospitals. In the health plan context, as noted above, the regulator could be an employer, a coalition of employers, or the government.

[33] In Shleifer's model pricing at average cost is second best; production is efficient, but there is the standard deadweight loss from prices above marginal cost. Because health insurance often breaks the link between demand and supply prices, the deadweight loss from average cost pricing may not matter in the medical case.

V. *Relaxing Shleifer's Assumptions: "Legitimate" Cost Heterogeneity*

By relaxing three of Shleifer's assumptions, the health economics literature reaches a different conclusion; reimbursement should not necessarily be fully prospective, though the optimal degree of prospectivity remains open. This section and the subsequent two take up the three assumptions in turn.

Some literature relaxes Shleifer's assumption of a homogeneous product by focusing on unobserved heterogeneity across consumers—they are differentially costly to treat—or sometimes observed but unreimbursed heterogeneity (Gregory Pope 1990; Colin Goodall 1990; Emmett Keeler 1990; Thomas Selden 1990; Ellis 1993; Ma 1994). The connection to the heterogeneity in the selection literature, however, is generally not made; Ma (1994) is an exception. Because of unobserved (by the regulator) heterogeneity across consumers, much of this literature suggests a mixed fee-for-service and lump-sum method of reimbursement; the lump sum may be at the level of the patient-year (capitation) or the admission.

Pope (1990) illustrates both the method and a problem of this literature. He assumes a simple hospital cost function:

$$C_i = T_i + s_i + e_i, \qquad (1)$$

where C_i is cost at the i^{th} hospital, and T, s, and e are components of cost arising respectively from quality, unobserved patient heterogeneity across hospitals, and managerial effort; e includes slack. The regulator wishes to reimburse only T

and s. Because the regulator wishes to reimburse s, Shleifer's solution of reimbursing the average cost of all other hospitals underpays hospitals with high values of s. Importantly, the link with the selection literature is severed because s is assumed exogenous.

Pope postulates that the regulator wishes to minimize squared error loss in payment, where payment to the i^{th} hospital, P_i, is a linear function of observed cost C_i:

$$P_i = R + \alpha C_i, \tag{2}$$

and the desired payment P^d is:

$$P^d = E(T) + s_i, \tag{3}$$

$E(T)$ is the cost of an average or desired quality level. $\alpha = 0$ is fully prospective payment; $R = 0$ and $\alpha = 1$ is cost reimbursement.

Minimizing mean square error implies the regulator chooses R and α to minimize $(1/n)\Sigma(P^d - P_i)^2$. Substituting for P^d and P_i in this expression and rearranging, one obtains:

$$(1/n)\Sigma(s_i - (R - E(T)) - \alpha C_i)^2, \tag{4}$$

which is the least squares formula for a regression of s, the unobserved cost component, on C, observed cost, and an intercept term, $(R–E(T))$. Although s is not observed, Keeler (1990) shows that if R were given by the *PPS*, α should be at least 0.15 to 0.29.

Pope's result poses an important problem: the payment formula $R + \alpha C_i$ leaves some hospitals incurring losses unless R is sufficiently high.[35] Suppose, in the spirit of yardstick competition, that R is set at the average cost of all hospitals, with hospitals reimbursed a portion $\alpha < 1$ of any overage. Then all hospitals with above average costs will lose money, albeit less than under fully prospective reimbursement at the average cost.

This problem is addressed by Jean-Jacques Laffont (1987) and Laffont and Jean Tirole (1986, 1993). Their context is the related problem of government procurement and in particular the use of fully prospective ("high-powered") vs. cost-reimbursement or cost-plus contracts. Cost-reimbursement contracts are not the same as fee-for-service, but they resemble each other in that both are low-powered, giving little or no incentive to economize on the quantity of services if fees are at or exceed the first-best level C^*.

Laffont and Tirole relax several of Shleifer's assumptions:

1) Like the health economics literature, firms are not assumed to be homogeneous with respect to cost. Although observed total cost varies with managerial effort, as in both Shleifer and Pope, Laffont and Tirole assume observed cost is also a function of a cost parameter that varies by firm and is unknown to the regulator, analogous to Pope's s_i. Thus, a regulator cannot tell from the realized costs whether the manager is lazy or whether the firm is simply high cost. A key assumption, however, is that the regulator knows the firm's true cost parameter lies between certain limits.

2) The regulator wishes to keep several firms in business, not just the lowest cost firm(s). This is also analogous to Pope's model where the regulator will pay for s_i. Whereas Shleifer's model is in the spirit of perfect competition—firms can costlessly enter—entry is not possible here; the regulator deals with existing firms, whatever their underlying

[34] Assuming that the regulator wished to reimburse e at the optimal level would not change the spirit of Pope's results.

[35] Pope's model also requires two strong assumptions if minimum squared-error loss is to result in an efficient outcome: Patients have identical marginal valuation curves that linearly decline in T; and $T - T^*$, the deviation from the desired intensity, must be proportional to $P - P^d$, the deviation from the desired price.

costs, and intends to keep at least a subset of them in business.

3) Revenues to pay firms are raised through a process with deadweight loss, typically taxation, although employer-paid health insurance can also cause deadweight loss. To minimize deadweight loss, the regulator tries to minimize rents firms receive.

This last assumption means that the maximal high-powered incentive, a lump-sum payment set at the level of the highest cost firm the regulator wishes to keep in business, is generally not optimal. It leaves lower cost firms earning rents, which must be paid for through taxes or some other distortive mechanism. In effect, the regulator trades off the welfare loss from raising revenues to pay rents with the welfare loss from the production inefficiency from non-fully prospective payment.

4) Unlike the Pope model, the firm must break even.

Laffont and Tirole's context is that of firms bidding for a contract. The regulator wishes to offer lower cost firms an incentive to reveal their true costs by submitting low bids. This is achieved through a reimbursement mechanism that is a linear combination of a lump sum and observed cost, with the weight on the lump sum increasing as the bid is lowered. At the low cost extreme, the firm keeps any residual. At the high cost extreme, it keeps none; the contract is cost reimbursement.

Laffont and Tirole's model could possibly be adapted to the health care context. The underlying cost parameter could, for example, arise from unobserved case mix variation, Pope's s_i, which a regulator might wish to reimburse. Bidding, however, fits more naturally with defense procurement than with health care if consumers are allowed choice. In that event bidders, for example HMOs submitting bids to care for a firm's employees, do not necessarily know which patients will choose them when they submit a bid.[37] Thus, unlike the defense case, health care firms do not necessarily know their true cost parameter when bidding.

Nonetheless, one could imagine each HMOs bidding a price schedule, a fixed amount for each enrollee and a variable amount depending on each enrollee's use. Selecting winning bids requires a method to combine these two pieces of information; how to do so poses difficulties that I cannot explore here.

Laffont and Tirole derive a closed form solution to their model. Unfortunately that solution requires more information than is usually available: lower and upper bounds of the distribution of true cost across efficient firms; the function relating a firm's cost to managerial effort; and the distortionary cost of finance.

Ma (1994) describes a hospital model that makes many of the above assumptions: patients vary in their costliness; they can be treated with varying intensity and demand responds to that variation; and enhancing quality or re-

[36] The standard deadweight loss from employer-paid premiums is unemployment among minimum wage workers to whom a premium cannot be passed on (Bridger Mitchell and Charles Phelps 1976). A premium will also distort labor supply if health benefits at the margin are not valued at their cost and if labor supply has some elasticity (Lawrence Summers 1989). Moreover, if the premium were to be mandated, it would almost certainly be accompanied by tax-financed subsidies to low-wage employees or small firms. Such subsidies would create the standard deadweight loss.

[37] Firms move first and submit prices; then consumers choose.

[38] The latter component might be a conversion factor for an exogenously given set of relative prices (a fee schedule).

[39] Moreover, how to calculate how much of a plan's bid should be passed on to consumers as a premium difference is an open question; the intent would be to pass on differences representing inefficiency or style of practice but not differences in risk mix.

ducing cost requires managerial effort that must be compensated. Unlike Laffont and Tirole, he does not consider deadweight loss from financing rents.

After showing that fully prospective reimbursement elicits the efficient level of managerial effort if all patients must be treated (the regulator in effect chooses the payment rate P^* that achieves the desired intensity level at each hospital), Ma considers selection. He shows that a piecewise linear scheme analogous to reinsurance is optimal; the hospital is paid P^* for patients whose costs are less than P^*, the profitable patients for whom dumping is not an issue, and paid cost for other patients. Because the hospital earns rents on the low-cost patients, however, the optimality of Ma's result depends on the assumption of no deadweight losses in financing the rents.

VI. *Relaxing Shleifer's Assumptions: Agency*

All the above models assumed a unitary firm that either maximized profits or whose managers maximized their utility. Ellis and McGuire (1986), working with the example of reimbursing hospitals, modify this assumption by introducing the physician as a utility-maximizing agent for both the hospital and the patient. Although developed in the context of hospitals, one could also apply the model to physicians working for a HMO. Ellis and McGuire conclude that fully prospective payment is only optimal in a special case of the physician's utility function.

In their model patients benefit from a greater quantity of costly hospital services. If the hospital is reimbursed prospectively, the physician, when deciding to order an additional service, must trade off patient benefit and hospital profit. The original Ellis and McGuire model assumes the physician controls this tradeoff (the patient is passive); a subsequent model (1990) generalizes to bargaining between the physician and the patient, but yields the same qualitative solution.

Given the physician's utility function with hospital profit and patient benefit as arguments, an efficient solution implies the marginal dollar of patient benefit costs a dollar of hospital profit. Thus, to solve the model requires knowing the marginal rate of substitution in the physician's utility function between patient benefit and hospital profit; if the physician trades them off equally, fully prospective payment is optimal, but if the physician gives greater weight to hospital profit, the reimbursement scheme can compensate by moving toward cost reimbursement, or a less high powered reimbursement scheme.

The same general argument applies to incentives facing physicians to over- or underservice if fees do not equal C^*. If pure fee-for-service results in overservicing and pure capitation in underservicing, the optimal scheme is a mixture.

VII. *Relaxing Shleifer's Assumptions: Errors in Pricing*

To this point I have asked the question: Suppose a regulator or insurer sets supply prices; should those prices be at a

[40] Ma also shows that fully prospective payment is optimal if the hospital can manipulate intensity within a payment class so as to perfectly discriminate by costliness of patient; that is, by varying intensity, the quantity demanded from patients who differ in their costliness of treatment is first best for each class of patients. This result, however, also leaves the hospital earning rents. The practical significance of Ma's findings is problematic. The result on quality discrimination does not seem easy to implement if patients are more or less continuously distributed with respect to cost, as in fact they often are. Further, explicit discrimination by patient can run counter to medical ethics, and many inputs in a hospital setting appear to be local public goods; for example, the number of nurses at a nurse station.

finely disaggregated fee-for-service level, a partially aggregated level (e.g., a set price for all services in a hospital admission, as in the Medicare PPS), or a highly aggregated level (capitation for all services during a given time period)? I have also observed that a regulator or insurer is likely to err and not set prices equal to the first-best C^*. Errors arise because the regulator will be uncertain about the cost function, as in the Laffont-Tirole model, and from technological change and learning-by-doing, which cause regulatory lag.

Newhouse (1991) pointed out that the resulting deviation from first-best pricing creates welfare losses. One likely loss in a fee-for-service world was described by Pauly; rents could induce too many services. Alternatively, rents in either a fee-for-service or a capitation world could be wastefully competed away by increased intensity. Because welfare loss is a function of squared errors, Newhouse argued that a linear combination of bases of payment averaged pricing errors and therefore reduced welfare loss. For linear demand and supply curves and unbiased payment systems (i.e., expected pricing errors of zero), the optimal combination of bases is linear, with weights inversely proportional to the error variances in the bases.[42]

VIII. *Will a Competitive Market Achieve Optimal Prospectivity?*

In the tradeoff between moral hazard and risk aversion, one could expect a competitive insurance market to achieve the optimal amount of demand-side cost sharing. Such a conclusion does not necessarily carry over to the amount of supply-side cost sharing.

In Jacob Glazer and McGuire's (1994) model hospitals treat patients of two payers (insurers) using both contractible (e.g., the operating room) and noncontractible inputs (e.g., quality of the nursing staff). The latter inputs have fixed costs. Each payer has an incentive to shift those fixed costs to the other payer(s) and may be able to do so by the amount of supply-side cost sharing employed in its choice of reimbursement contract. For example, a prospective payer facing a cost-based payer can cause the hospital to shift fixed costs to the cost-based payer, a result similar to Danzon's (1982). In the Glazer-McGuire model the competitive (symmetric) equilibrium may be either too prospective or insufficiently prospective to achieve production efficiency.

Glazer and McGuire find a range of optimal solutions that use fully prospective payment.[43] Thus, in the context of their model, the optimality of a corner solution is more plausible than with demand-side cost sharing. Importantly, however, their model does not consider selection behavior.

IX. *Should Payment Systems Be Mixed?*

The foregoing models consider payment methods that elicit socially efficient production, or as efficient as possible given information asymmetries. They suggest that Shleifer's conclusion that

[41] As was the case with airlines competing away rents under regulated fares (Clifford Winston 1993). Philip Held and Pauly (1983) and Paul Joskow (1983) make an analogous argument for Medicare payments for renal dialysis.

[42] This follows from the formula for the variance of a mean and assumes independent errors.

[43] In their model, as in the Ellis-McGuire model, hospitals' utility functions are defined over patient benefit and profit, and the greater the weight on profit, the less likely it is that fully prospective payment is optimal. Also the larger the fixed costs, the less likely fully prospective payment is optimal.

[44] In fairness to them, selection behavior may be less important in the hospital context that they consider than in the insurance plan context; certainly the empirical evidence for selection in the hospital context is sparser than in the insurance plan context, as described below.

fully prospective payment is optimal is not robust against more general assumptions. Save for Ma, however, none of the models formally considers selection behavior.

Accounting for selection, however, strengthens the conclusion of these models that the corner solution of full prospectivity is not likely to be optimal because the welfare loss increases with any additional selection. Even apart from the models just reviewed, however, one might argue for a mixed payment method. Under standard assumptions the marginal welfare loss from insufficient managerial effort increases as the degree of prospectivity is reduced, and I noted in Section III that the marginal welfare loss from selection as the degree of prospectivity increases may increase over some range. Under these conditions the tradeoff between efficiency in production and selection will be (locally) strictly concave to the origin, suggesting an interior (mixed reimbursement) rather than a corner solution.

Even granting that payment should be mixed, most models in the literature consider only linear combinations of different bases of payment. But even linear models do not seem ready to apply because optimal weights require information on unobserved quantities: Laffont and Tirole's solution requires information on the manager's utility function for effort; Pope's requires information on unobserved patient factors affecting cost; Ellis and McGuire's requires knowing the physician's utility function; and Newhouse's requires information on error variances in the prices.

[45] Ellis and McGuire (1986) remark that prospective payment of hospitals may cause selection, but that is outside their model.

[46] Nonlinear models are observed in practice; see below.

[47] A similar result on mixed reimbursement also appears in a class of models from estimation theory (Carole Siegel et al. 1992), but these models reverse the usual assumption by assuming the regulator wants to reimburse institution-specific costs and that those costs are invariant to managerial effort. Nonetheless, the regulator does better in a mean square error sense to reimburse a linear combination of institution-specific costs and an average cost over all institutions rather than institution-specific costs.

[48] If marginal cost is less than average cost, of course, marginal cost pricing requires lump-sum transfers.

[49] Rogerson (1990) makes the same assumption in the context of defense contracting; the Department of Defense is analogous to the hospital and Congress to the regulator.

[50] Rogerson performs the following thought experiment: Suppose a hospital could employ two intensities. The regulator will reimburse the cost of either; assuming the costlier intensity provides more gross benefit, the hospital prefers it. The

X. *A Refinement: Varying Price-Cost Margins within a Payment Base*

To this point I have focused on the basis of price—fee-for-service or capitation—rather than how price was set within any basis. Within any basis I have presumed that, if feasible, price should be set at C^*, the marginal cost associated with the competitive quantity and quality of service. William Rogerson (1994) shows this presumption is false; a regulator should not necessarily set price at C^* even if it is feasible.

Rogerson points out that prospective payment for a hospital admission not only gives the hospital an incentive to produce efficiently but also gives the regulator a method to control the intensity of production. He concentrates on how the regulator and hospital choose intensity.

Rogerson assumes hospitals are nonprofit and therefore choose intensity to maximize *gross* social surplus; the regulator, however, wishes to maximize *net* social surplus (gross surplus less cost), subject to a breakeven constraint.[49] Therefore, in choosing how to provide care the hospital prefers more intensive treatment, and under cost reimbursement intensity will be too high. Put an-

other way, Rogerson assumes the hospital attempts to obtain the maximum reimbursement and thus produces with maximum intensity.

If there is one product (i.e., treatment of one disease) with multiple intensities (e.g., varying levels of nurse staffing), the regulator can choose a price to reach the desired intensity. In this case price equals C^*, assuming for simplicity that marginal cost pricing satisfies the break-even condition. With multiple products (diseases), however, the analysis differs markedly. If one generalized the result from the single product case, the regulator would set a price such that the cost of the desired intensity for each product was exactly reimbursed. This appears to be the intent of both Medicare's PPS for hospitals and its Resource Based Relative Value Scale for physicians, both of which attempt to set prices at the average cost of treating a given disease.[51]

In fact, such a policy does not lead to the desired intensities because of hospital behavior. Suppose there are two products, and the regulator sets prices such that the hospital breaks even if it produces each product at the regulator's desired intensity. Rogerson assumes that demand at the hospital responds to its treatment intensity. If the intensity elasticities for the two products differ, the hospital can reduce intensity for the product with the low-intensity elasticity and increase it for the product with the high-intensity elasticity, thereby increasing its reimbursement and increasing gross social benefit.

To counteract this tendency, the regulator allows varying price-cost margins at the desired intensity level. Analogous to Ramsey pricing, the margin is greater for the product with the low-intensity elasticity. In equilibrium, hospitals earn positive accounting profits on products with low-intensity elasticities and negative profits on products with high-intensity elasticities.[52]

Rogerson's model of hospital pricing resembles Gerald Wedig's (1993) model of physician pricing. Wedig assumed that average cost exceeded marginal cost, so fees would have to exceed marginal cost for a physician to break even, and that physicians could exploit the resulting price-cost margin to induce demand. The degree of inducement, however, could vary by service, if the cost of inducement varies, e.g., low for a noninvasive diagnostic test; high for a surgical procedure with a nontrivial risk of operative mortality. As a result, a regulator varies the markup across services according to the supply elasticity of the service.

Varying margins within a payment base can be employed with payment methods that use combined bases. In other words, with the knowledge of the requisite elasticities, a regulator could alter a partial prospective payment method and improve over a partial system that was not so altered.

XI. *Regulation of the Selection-Efficiency Tradeoff*

Any pricing method is set in an institutional context, and one may ask whether altering that context through non-price regulation might improve the selection-efficiency tradeoff. The selection problems that arise in health care have analogues elsewhere, and regulation has sometimes tried to enforce a pooling equilibrium. For example, before airline deregulation the Civil Aero-

regulator, however, cares about gross benefits less cost, which may or may not be maximized by choosing the costlier intensity.

[51] I ignore the problem of allocating joint costs, which is not the issue here.

[52] In both Jeffrey Harris' (1979) and Danzon's (1982) models of hospital behavior, price-cost margins differ for various hospital services or to various payers because hospitals are assumed to be discriminating monopolists.

nautics Board required airlines to serve unprofitable routes and cross subsidize them from profitable ones. Similarly, long distance rates subsidized local telephone service.

The health care analogue would require a health plan to treat patients whose expected costs exceed the capitation reimbursement, with the loss offset by profits on other patients. Analysts differ on how effective regulation could be in enforcing nondiscrimination among patients, i.e., in enforcing a pooling equilibrium, but I am pessimistic.

One context for anti-selection regulation is an employer who offers a choice of plans. In this context existing anti-selection regulations include periodic open enrollment (at a specified time in the year anyone in the group can enroll in any plan without medical underwriting) and guaranteed renewal (anyone in a plan can remain in a plan at the same terms available to others).

Such regulations, though presumably enforceable, do not necessarily suffice to prevent selection because insurance plans, especially integrated ones, can alter their product to influence choice. They may selectively market, choosing media seen mainly by better risks. They may site facilities to appeal to better risks. Their staffing may discourage some types of risks and encourage others; for example, they may stint on oncologists (cancer specialists) but have numerous pediatricians (families with children are better risks).[53] Staffing choices seem especially hard to regulate, because of numerous sensible opportunities for substituting less highly trained personnel for specialists.[54] Plans can also try to alter the patient encounter. They could offer incentives to gatekeeper physicians not to refer patients to specialists, thereby discouraging enrollment by the chronically ill who wish to see a specialist. They could pass the capitation for a patient through to the gatekeeper physician or small group of physicians (i.e., make them partial residual claimants), thereby causing the physician to alter behavior in subtle ways so as to convince the unprofitable patient to seek another plan.[55]

Consumer behavior may also undercut regulation. New plans with new networks, for example, will appeal more to good risks, because the chronically ill may have a set of physicians, some of whom may be outside the new plan's network. As a result, the lower the barriers to plan entry, the greater the selection potential.

XII. *Empirical Literature on Selection and Plan Reimbursement*

One empirical literature enquires into the degree of selection behavior actually observed. Another aims at developing "risk adjusters" for premiums in order to minimize selection for any given degree of prospectivity; some of this latter literature attempts to define subpopulations within which payment could be fully prospective, and some of it abandons fully prospective payment. I discuss these two literatures in turn.

Empirical Evidence on Selection. One

[53] One device aimed at countering selection is a so-called carve-out benefit; a payer such as an employer in effect mandates that a certain fraction of the premium dollar be spent for specific benefits such as mental health or cancer benefits. This can be done by competitively contracting for these benefits separately. Separate contracts may, however, result in coordination problems.

[54] Note that the integrated insurer-deliverer-of-care is better situated to engage in such tactics than the traditional nonintegrated insurer whose plans offered free choice of provider.

[55] To counter such actions the organizer of the insurance menu, for example the employer, could in principle threaten to exclude from the menu plans who were cream skimming. But detection is problematic, and exclusion from the menu could well be resisted by the majority in the plan who benefit (at least in the short run) from the cream skimming.

TABLE 2
PRIOR USE OF MEDICARE SERVICES BY HMO AND NON-HMO ENROLLEES, ADJUSTED FOR DEMOGRAPHICS, 1985–1986

Ratio of Prior Medical Expense, HMO Enrollees to non-HMO Enrollees	Percentage of the 98 plans in interval
<0.80	56
0.80–1.00	38
1.01–1.20	6
Average ratio among all 98 plans, 0.77	

Source: Hill and Brown (1990, table III.7)

TABLE 3
POST-ENROLLMENT ADJUSTED MORTALITY RATES, HMO AND NON-HMO ENROLLEES

Ratio of Adjusted Mortality Rates Among HMO Enrollees to Non-HMO Enrollees in the post-enrollment period	Percentage of Plans
<0.80	58
0.80–1.00	42
Overall average ratio among 83 plans, 0.75	

Source: Hill and Brown (1990, tables III.19 and III.20).

of the most convincing studies that plans and insurers either actively or passively exploit uniform pricing in the face of risk heterogeneity is Jerrold Hill and Randall Brown (1990). They studied choices of Medicare beneficiaries who were offered the option of enrolling in an HMO or remaining in traditional, fee-for-service Medicare. Hill and Brown studied 98 HMOs, comparing the groups who chose the HMO with those who did not. In the year prior to the choice, when both groups were in the traditional plan, those who subsequently chose the HMO spent 23 percent less than the others, after adjusting for differences between the groups with respect to age, sex, welfare status, institutional status, and location (county) (Table 2).

Hill and Brown also collected data on mortality in the year following the choice for 83 of the 98 HMOs. After adjustment, mortality was less in all 83 HMOs than in the comparison groups, on average 25 percent less (Table 3). Although better care in HMOs could account for this result, such an explanation seems much less probable than that healthier individuals chose the HMO option.

Medicare pays HMOs for each enrollee 95 percent of the average area fee-for-service Medicare payment (adjusted for age, sex, welfare status, institutional status, and county), so plans could profit by enrolling better risks; the evidence in Tables 2 and 3 demonstrates that these incentives mattered. Of course, Medicare could lower its payment to HMOs and recoup the plan profit. Given risk heterogeneity, however, there would still be a profitable subgroup, and the process could repeat. Further, the unprofitable group, for whom the competing plan market fails, would increase.[56]

Cutler (1994) showed that the average risk varied substantially among nonintegrated insurance plans (Table 4). A family policy at the 90th percentile of the premium distribution cost almost three times as much as one at the 10th percentile ($7670 vs. $2760). If the same population were given these two policies, however, predicted spending with the 90th percentile policy only exceeded that with the 10th by 40 percent ($5890 vs. $4220); the difference is risk heterogeneity. Whether this selection comes through poor risks' wanting more extensive insurance or insurers charging them more, the effect is a limitation on the ability to insure against becoming a bad risk in the future.

[56] Market failure in this context means the person enrolls in traditional Medicare.

TABLE 4
COST AND ACTUARIAL VALUE OF INSURANCE POLICIES

	Individual Policy		Family Policy	
Percentile	Cost	Actuarial Value	Cost	Actuarial Value
10	$1220	$1740	$2760	$4220
25	1670	1910	3950	4600
50	2100	2100	5070	5070
75	2620	2260	6090	5459
90	3220	2440	7670	5890
Difference				
90–10	164%	40%	178%	40%

Source: Cutler (1994, table 2).

A similar example comes from the Federal Employees Health Benefits Plan, which offered two nonintegrated plans with free choice of physician, known as high- and low-option, differing only in the amount of cost sharing. In 1982 the high-option premium was nearly double that of the low option, whereas the actuarial difference was less than ten percent (James Price, James Mays, and Gordon Trapnell 1983; Price and Mays 1985). The two plans were marketed similarly and differed only in the amount of cost sharing (e.g., the size of the deductible). The sorting came from the greater appeal of less cost sharing to bad risks.[57] Subsequent switching out of the high-option plan made the existence of an equilibrium with both these plans in the market problematic (W. Pete Welch 1989).

These studies give concrete examples of contracts that limit pooling. HMOs attracted good risks, as did less generous, nonintegrated plans. Some generous plans may be driven from the market, the so-called premium death spiral.

Risk selection at the level of the hospital admission has been less well studied than at the plan level and when studied, results have been more ambiguous. Newhouse (1989) found evidence of selection in the PPS, but it was quantitatively modest, perhaps because Medicare payments to hospitals at that time mostly exceeded average accounting cost and because hospitals had excess capacity.[58] Nonetheless, Newhouse and Daniel Byrne (1988) found that long stay patients, those who would have been costly under the PPS, were differentially admitted to facilities exempt from the System. And McClellan (1993) found a sharp fall in medical admissions and increase in surgical admissions after PPS, consistent with its pricing incentives.

Other literature documenting selection effects includes Sy Berki and Marie Ashcraft (1980), Berki et al. (1978), Ellis (1985, 1989), Fred Hellinger (1987, 1995), Harold Luft and Robert Miller (1988), McGuire (1981), and James Robinson, Laura Gardner, and Luft (1993). I now turn to studies aimed at

[57] An extreme version of selection was seen in high- and low-option plans offered Harvard employees that differed only in the amount of the deductible, but the premium difference exceeded the deductible! Nonetheless, a sixth of the group opted for the high-option plan in 1988.

[58] Occupancy rates have been around 60 percent in recent years.

modifying payment formulas to improve the selection-efficiency tradeoff.

Risk Adjustment. Risk adjustment is the term used to characterize the literature on adjusting premiums for enrollee characteristics. The growing price competition among vertically integrated insurance plans has stimulated this literature, which uses variables such as age and gender either to group reasonably homogeneous individuals or to adjust a plan's premium revenue. The intent is to match the enrollee's expected cost more closely and thereby reduce plan incentives to select good risks.

This literature typically uses explained variance in annual individual expenditure to judge the goodness of risk adjusters. It usually estimates an equation of the form:

$$Y_{it} = \alpha + X_{it}\beta + \mu_i + \varepsilon_{it}, \qquad (5)$$

where Y_{it} is the i^{th} person's medical care consumption in the t^{th} time period, α and β are constants, X_{it} is a vector of risk adjusters; μ_i is a time invariant, person specific effect with a mean of zero, and ε_{it} is a random error term with an expected value of zero.[59] In the case of pure capitation there are no X's, and the premium is simply α, the mean spending for the insured group, plus a loading. Thus, the search for risk adjusters attempts to minimize the variance of μ and to remove autocorrelation from ε.

Of course, it is not total variance but only the predictable portion that needs explanation. The predictable portion is the between-person variance, plus any predictable portion of the within-person variance.[60] If families enroll as a unit, they become the relevant unit of observation, and the predictable fraction of variance will increase because random events will be averaged across family members.[61]

Fixed effects models for annual expenditure estimate the between-person variance to be around 15 to 20 percent of the total (Nelda McCall and H. S. Wai 1983; Newhouse et al. 1989). Additionally, some within-person variance is predictable because some time varying covariates are predictable—the simplest is age—and because the error term has an autoregressive component. After adjustment for several observables, the $AR(1)$ correlation coefficient appears to be on the order of 0.2.[62] Thus, one should add to the 15 to 20 percent figure another four percentage points, making the predictable portion, exclusive of time-varying covariates, around 20 to 25 percent of total variance. Total explained variance is only a first-order approximation to a criterion of goodness, because one is also concerned about explanation within subgroups or, in the case of a continuous adjuster, the fit of the function over the entire range of the adjuster. Although adding a risk class should increase total explained variance, it may decrease it in certain subclasses, thereby increasing the incentive for selection in those subclasses.

---

[59] One could of course include a fixed or random effect that varied with time; for economy of notation I have omitted it.

[60] More precisely, it is $\alpha + X_{it}\beta + \mu_i$ + any predictable amount from observing past values of Y_i. Time periods longer than a year cause the predictable fraction of total variance to increase, because expenditures arising from random events will be averaged to a greater degree. Actual enrollment periods, however, rarely exceed a year and are sometimes shorter; e.g., Medicare allows an enrollee to change plans monthly.

[61] Individual- or family-specific risk adjusters are unimportant if an entire group that is formed for purposes other than obtaining insurance (e.g., employees of a firm) is enrolled in a single plan. In that case, average prior utilization in the group should predict average future utilization well. This is experience rating.

[62] The value 0.2 comes from averaging correlations between residual spending in Table 3 of Newhouse et al. (1989) and from fitting an equation to data from James Beebe as described in the appendix to Newhouse et al. (1989).

How do existing adjusters stack up against the 20 to 25 percent criterion? Many analysts have quantified the contribution of various adjusters to explained variance (Arlene Ash et al. 1989; Arnold Epstein and Edward Cumella 1988; Jonathan Howland et al. 1987; Newhouse et al. 1989, 1993; Helen Schauffler, Howland, and Janet Cobb 1992; Jonathan Weiner et al. 1991). In general, demographic characteristics such as age and sex, which are inexpensive to collect and easy to audit, explain little variance, around one percent. Race explains an additional modest amount, but is a forbidden adjuster. Geography also explains a modest amount.[63]

Much of this literature tries to exploit individual health status information. The results have been mixed, but on the whole discouraging. Self-reported measures, either summary subjective measures or more objective measures of such traits as physical mobility, have two drawbacks: they leave substantial variance unexplained and pose potential auditing difficulties. Physiologic measures may indicate the cost of treating a chronic condition, but they can be costly to obtain. Moreover, if treatment affects the measure (e.g., medication to control blood pressure), one may lack a measure of the severity of the disease before treatment or its severity may change during treatment in ways that cannot be separated from the effect of the treatment.

Initial results on variance explained by chronic conditions were not encouraging; recent results are more positive. Brown et al. (1993) find that accounting for a history of heart disease, cancer, and stroke reduces residual explainable variance among the Medicare population to modest levels, but these measures may be susceptible to gaming. For example, the PPS system adjusts payment for a hospital admission based on diagnosis and certain procedures, because costs of treating different diagnoses vary. When it was instituted, a certain amount of upcoding (i.e., coding patients as having more severe disease than similar patients had previously been coded) took place in order to increase reimbursement (Grace Carter, Newhouse, and Daniel Relles 1990). This may simply illustrate that, to paraphrase Lord Acton, data may be corrupted, but data used for payment may be corrupted absolutely.

Typically the measures that have explained the most variance have been some function of prior utilization. Prior utilization, of course, is a lower-powered basis of payment and potentially reduces production efficiency; for individuals who remain in a plan, prior utilization as a risk adjuster differs from current utilization only by a discount factor.[64] Thus, a payment formula using prior utilization is similar to a mixed payment method.

Mixed payment systems exist. The PPS outlier system provides reinsurance at the patient level if a cost threshold is exceeded. The Diagnostic Cost Groups system (Ash et al. 1989) pays an HMO more in period t if a person is hospitalized for

[63] Although a competitive market will tend toward rating small geographic areas (e.g., auto insurance rates five-digit zip codes), the skewness of medical spending makes small-area means unstable. Because of that instability, the Medicare risk adjustment formula, which adjusts for geography at the level of the county, uses a five-year average of county to national spending to adjust an annual national average. A better method appears to be to shrink small-area means toward a larger area mean using Empirical Bayes methods (Herbert Weisberg, Thomas Tomberlin, and Sangit Chatterjee 1984; Tomberlin 1988; Noel Cressie 1991).

[64] For those who do not remain in a plan, there are competing considerations. Prior use offers the plan an incentive to keep a sick enrollee, especially if the disease causes a substantial short term expense but only a small increase in expected long term spending. Otherwise the plan will not recoup the current spending. On the other hand, there is a stronger incentive to invest in arrangements that reduce the number of chronically ill enrolling.

a relatively nondiscretionary purpose in period $t - 1$ and if that hospitalization predicts spending in t. This resembles Rogerson's and Wedig's suggestions to raise price-cost margins when supply is relatively inelastic.

There remains the question of how good adjusters have to be to render selection unimportant; can they be good enough to leave fully prospective payment optimal? Unfortunately the returns to a plan from exploiting private information to select are nonlinear; if a plan could costlessly discriminate, it needs to explain only a modest amount of residual variance to reap a substantial reward (Newhouse et al. 1989).[65] It follows that the formula for adjusting for heterogeneity must be close to perfect to reduce greatly the incentives to select. Present risk adjustment methods are a long way from perfection. It is not surprising that the studies cited above find evidence of selection.[66]

XIII. *Policy Implications*

Mixed Payment Systems and Insurance Plan Reimbursement. Plans typically distinguish between individuals and families in their quoted premiums (Table 4), but they often do not distinguish by age, sex, or even the number of children. Put another way, plans make little attempt to experience rate individuals or families. One can infer that to do so is costly, and that the costs are high enough to maintain a limited pooling equilibrium. Nonetheless, plans paid fully prospectively have clear incentives to restrict the scope of any such pooling. The empirical literature suggests that fully prospective payment results in strong incentives to select good risks and that such selection takes place.

Selection can cause markets to be incomplete. Its extreme is a form of redlining; no supplier wants the bad risk (or bad risk within a class). In its less extreme versions an individual cannot insure against the higher premiums from becoming a bad risk or cannot obtain as much insurance as desired because of asymmetric information. Further, because a plan's price reflects the average risk of its enrollees and the market segments by risk, price differences to persons choosing among plans reflect more than style and efficiency differences, thereby potentially causing persons to misallocate themselves among plans.

Mixed payment systems reduce the incentive to select, with correspondingly less incentive to produce efficiently. The literature contains several models of mixed systems, but they are mainly linear and the optimal weights they imply require knowledge that is unavailable or strong and untestable assumptions or both.

For that reason determining preferred weights will likely require experimentation. Both linear and nonlinear schemes should be considered. For example, the optimal weight on actual use might rise with use, as with reinsurance above a threshold. In any event, a regulator or employer cannot sensibly decide about the degree of prospectivity without knowing the terms of the tradeoff between efficiency in production and selection.[67]

Mixed Systems and Hospital and Physician Reimbursement. The conclusion from the literature on hospital and physician reimbursement is the same as that

[65] As adjusters improve in predictive power, the degree of nonlinearity falls (Wynand van de Ven et al. 1994).

[66] On the other hand, simulations by Susan Marquis (1992) suggest the use of age and sex as adjusters could prevent a premium death spiral of the type suggested by the Federal Employees experience.

[67] For lack of any information I have not considered the additional cost of administering a mixed reimbursement system.

on plan reimbursement; fully prospective payment is probably not optimal. In the case of hospital reimbursement it seems plausible that the weight on the prospective payment could be higher than for an integrated insurance plan because of less scope for selection. First, physicians not hospitals admit patients, and under fee-for-service their incentives favor admitting patients who require more services.[68] Second, within Medicare's PPS, by far the largest example of prospective hospital payment, transfers are monitored and "dumping" is illegal.

Nonetheless, a hospital can select. For example, through its capital investments it can configure itself to select for (against) certain (un)profitable treatments or patients (McClellan 1993). That selection behavior is a concern is shown not only by the Medicare regulations that proscribe dumping but also by its outlier payment system, a move away from full prospectivity (Keeler, Carter, and Sally Trude 1988; Ellis and McGuire 1988).

Optimal physician reimbursement has attracted less attention, but use of full capitation is modest. United Kingdom general practitioners are capitated for their patient lists, with resulting incentives to over-refer to salaried hospital specialists, but there has been little investigation of whether these incentives matter. To increase incentives to produce efficiently, the National Health Service has recently made groups of 20 or more general practitioners (fundholders) financially responsible for all services used by patients on their lists. This in turn has raised selection issues (Manos Matsaganis and Howard Glennerster 1994).

For physicians under contract with or employed by a vertically integrated insurer, salary could be an additional basis of payment. In that case one needs to know what behavior is rewarded with raises, as well as the conditions of contract renewal.

Subsidies to Insurance Plans. To encourage pooling, most employers subsidize the price of a health plan to the employee. Although many employers traditionally paid more toward higher priced plans if they offered multiple plans, one competitive "reform" has been to move toward a uniform employer payment across plans. To the degree premium differences among plans stem from differences in risk mix, however, the traditional nonuniform payment may have actually been more efficient, given that more expensive patients chose more generous plans (Table 4). In other words, the traditional nonuniform payment could be treated as a crude form of risk adjustment.

Are Supply- and Demand-Side Cost Sharing Substitutes? Ellis and McGuire (1986) suggest that supply-side cost sharing might substitute for demand-side cost sharing, improving the terms of trade between moral hazard and risk aversion. The evidence, however, suggests the two types of cost sharing operate on different margins. Demand-side cost sharing affects consumer decisions to seek care, but has little influence on the cost of care once sought (Newhouse and the Insurance Experiment Group 1993, ch. 4). Supply-side cost sharing at best indirectly influences patient decisions to seek care; it aims at provider choices on treatment intensity for a patient who has sought care. Thus, optimality may well require both supply- and demand-side cost sharing.

Any-Willing-Provider Laws and Economic Credentialling. The emerging advocacy for several different laws and regulations suggests the importance of a form of selection, "economic credentiall-

[68] Different bases of payment for physicians and hospitals create different threat points in a bargaining game.

ing," or a plan's choosing physicians for its network on the basis of profit generated for the plan, thereby inducing the physician to discriminate against high risks. Among the most prominent proposals addressed at economic credentialling are any-willing-provider laws, which some states have adopted; they require certain integrated insurance plans to accept into their networks any provider willing to meet their price terms. Integrated plans argue that such laws hamper their choice of efficient providers and increase their monitoring costs. Analogous proposals include: requiring integrated health plans to publish their criteria for selecting or deselecting physicians; requiring due process in deselection; and requiring plans to offer some reimbursement if a patient sees a physician not in the plan's network.[69]

These proposals arise in part because of incomplete risk adjustment at the plan level; adequate adjustment leaves plans no incentive not to contract with physicians who treat costly patients. Economic credentialling concerns have also arisen around hospital staff appointments, suggesting that fully prospective hospital payment insufficiently adjusts for risk.

Quality Report Cards. Efforts to rate health plans, hospitals, or physicians with respect to quality, as for example with the Health Employer Data Information Set (HEDIS), may cause analogous selection problems. To the degree that such efforts use outcome-based measures that are incompletely adjusted for patient characteristics, plans have incentives to select patients who will make them look good. For example, if a plan knows Jones will not take his prescribed medications and the plan is being judged on Jones' outcomes, it may not want Jones as a member. The issue is the same as skimming applicants in manpower training programs when those operating the program are rewarded using observed outcomes. Report cards also give plans incentives to allocate resources toward improving performance on measured characteristics and away from activities affecting unmeasured characteristics. The alternative of no information, however, certainly lacks appeal.

Concluding Comment. In many, if not most markets the level of price is critical but its basis is not an issue. For example, it does not matter whether the price of apples is per pound or per ton. For medical services, however, not only the level of price but also its basis matters, both for which persons have what insurance coverage and how they are likely to be treated when sick. In particular, the degree of supply-side cost sharing affects the tradeoff between the incentives of insurers and providers to engage in selection and to produce efficiently; optimizing that tradeoff seems as important as optimizing the demand-side tradeoff between moral hazard and risk aversion.

[69] The American Medical Association, for example, advocates many of these provisions in the Patient Protection Act it sponsors.

References

AARON, HENRY J. "Issues Every Plan to Reform Health Care Financing Must Confront," *J. Econ. Perspectives*, Summer 1994, *8*(3), pp. 31–43.

AKERLOF, GEORGE. "The Market for 'Lemons'," *Quart. J. Econ.*, Aug. 1970, *84*(3), pp. 488–500.

ARROW, KENNETH J. "Uncertainty and the Welfare Economics of Medical Care," *Amer. Econ. Rev.*, Dec. 1963, *53*(5), pp. 941–73.

———. "The Economics of Moral Hazard: Further Comment," *Amer. Econ. Rev.*, June 1968, *58*(3), pp. 537–39.

ASH, ARLENE ET AL. "Adjusting Medicare Capitation Payments Using Prior Hospitalization Data," *Health Care Financing Rev.*, Summer 1989, *10*(4), pp. 17–29.

BARR, NICHOLAS. "Economic Theory and the Welfare State," *J. Econ. Lit.*, June 1992, *30*(2), pp. 741–803.

BAUMGARDNER, JAMES. "The Interaction Between Forms of Insurance Contract and Types of Technical Change in Medical Care," *RAND J. Econ.*, Spring 1991, *22*(1), pp. 36–53.

BERKI, SY AND ASHCRAFT, MARIE. "HMO Enroll-

ment: Who Joins What and Why: A Review of the Literature," *Milbank Memorial Fund Quart.*, 1980, *58*(4), pp. 588–632.

BERKI, SY ET AL."Enrollment Choices in Different Types of HMOs: A Multivariate Analysis," *Medical Care*, Aug. 1978, *16*(8), pp. 682–97.

BROOK, ROBERT H. "Maintaining Hospital Quality: The Need for International Co-operation," *J. Amer. Medical Assoc.*, Aug. 25, 1993, *270*(8), pp. 985–87.

BROOK, ROBERT H. AND MCGLYNN, ELIZABETH. "Maintaining Quality of Care," in *Health services research*. Ed.: ELI GINZBERG. Cambridge: Harvard U. Press, 1991, pp. 284–314.

BROWN, RANDALL S. ET AL. "Do Health Maintenance Organizations Work for Medicare?" *Health Care Financing Rev.*, Fall 1993, *15*(1), pp. 7–23.

CANTOR, JOEL C.; LONG, STEPHEN H. AND MARQUIS, M. SUSAN. "Private Employment–Based Health Insurance in Ten States," *Health Affairs*, Summer 1995, *14*(2), pp. 199–211.

CARTER, GRACE; NEWHOUSE, JOSEPH P. AND RELLES, DANIEL. "How Much Change in the Case Mix Index Is DRG Creep?" *J. Health Econ.*, 1990, *9*(4), pp. 411–28.

CHASSIN, MARK ET AL. "Does Inappropriate Use Explain Geographic Variations in the Use of Health Care Services? A Study of Three Procedures," *J. Amer. Medical Assoc.*, Nov. 13, 1987, *258*(18), pp. 2533–37.

CRESSIE, NOEL A. C. *Statistics for spatial data*. New York: John Wiley & Sons, 1991.

CUTLER, DAVID M. "A Guide to Health Care Reform," *J. Econ. Perspectives*, Summer 1994, *8*(3), pp. 13–29.

DANZON, PATRICIA M. "Hospital 'Profits': The Effects of Reimbursement Policies," *J. Health Econ.*, May 1982, *1*(1), pp. 29–52.

DARBY, MICHAEL AND KARNI, EDI. "Free Competition and the Optimal Extent of Fraud," *J. Law Econ.*, Apr. 1973, *16*(1), pp. 67–88.

DIAMOND, PETER. "Organizing the Health Insurance Market," *Econometrica*, Nov. 1992, *60*(6), pp. 1233–54.

DRANOVE, DAVID. "Demand Inducement and the Physician/Patient Relationship," *Econ. Inquiry*, Apr. 1988, *26*(2), pp. 281–98.

EGGERS, PAUL. "Risk Differentials between Medicare Beneficiaries Enrolled and Not Enrolled in an HMO," *Health Care Financing Rev.*, Winter 1980, *1*(1), pp. 91–99.

EHRLICH, ISSAC AND BECKER, GARY S. "Market Insurance, Self-Insurance, and Self-Protection," *J. Polit. Econ.*, July/Aug. 1972, *80*(4), pp. 623–48.

ELLIS, RANDALL P. "The Effect of Prior-Year Health Expenditures on Health Coverage Plan Choice," in RICHARD M. SCHEFFLER AND LOUIS F. ROSSITER, eds. 1985, pp. 149–70.

———. "Employee Choice of Health Insurance," *Rev. Econ. Statist.*, May 1989, *71*(2), pp. 215–23.

———. "Creaming, Skimping, and Dumping: Provider Competition for Patients." Mimeo, July 1993.

ELLIS, RANDALL P. AND MCGUIRE, THOMAS G. "Provider Behavior Under Prospective Reimbursement: Cost Sharing and Supply," *J. Health Econ.*, June 1986, *5*(2), pp. 129–51.

———. "Insurance Principles and the Design of Prospective Payment Systems," *J. Health Econ.*, Sept. 1988, *7*(3), pp. 215–37.

———. "Optimal Payment Systems for Health Services," *J. Health Econ.*, 1990, *9*(4), pp. 375–96.

———. "Supply-Side and Demand-Side Cost Sharing in Health Care," *J. Econ. Perspectives*, Fall 1993, *7*(4), pp. 135–51.

ENTHOVEN, ALAIN C. *Theory and practice of managed competition in health care finance*. Amsterdam: North-Holland, 1988.

EPSTEIN, ARNOLD M. AND CUMELLA, EDWARD J. "Capitation Payment: Using Predictors of Medical Utilization to Adjust Rates," *Health Care Financing Rev.*, Fall 1988, *10*(1), pp. 51–69.

GLAZER, JACOB AND MCGUIRE, THOMAS G. "Payer Competition and Cost Shifting in Health Care," *J. Economics and Management Strategy*, Spring 1994, *3*(1), pp. 71–92.

GOODALL, COLIN. "A Simple Objective Method for Determining a Percent Standard in Mixed Reimbursement Systems," *J. Health Econ.*, Nov. 1990, *9*(3), pp. 253–71.

GRUBER, JONATHAN. "The Incidence of Mandated Maternity Benefits," *Amer. Econ. Rev.*, June 1994, *84*(3), pp. 622–41.

HARRIS, JEFFREY E. "Pricing Rules for Hospitals," *Bell J. Econ.*, Spring 1979, *10*(1), pp. 224–43.

HELD, PHILIP J. AND PAULY, MARK V. "Competition and Efficiency in the End Stage Renal Disease Program," *J. Health Econ.*, Aug. 1983, *2*(2), pp. 95–118.

HELLINGER, FRED J. "Selection Bias in Health Maintenance Organizations: Analysis of Recent Evidence," *Health Care Financing Rev.*, Winter 1987, *9*(2), pp. 55–63.

———. "Selection Bias in HMOs and PPOs: A Review of the Evidence," *Inquiry*, Summer 1995, *32*(2), pp. 135–42.

HILL, JERROLD W. AND BROWN, RANDALL S. "Biased Selection in the TEFRA HMO/CMP Program," Princeton: Mathematica Policy Research, Inc., 1990 (MPR Reference No. 7786–503).

HOWLAND, JONATHAN ET AL. "Adjusting Capitation Using Chronic Disease Risk Factors: A Preliminary Study," *Health Care Financing Rev.*, Winter 1987, *9*(2), pp. 15–23.

JOSKOW, PAUL L. "Reimbursement Policy, Cost Containment, and Non-Price Competition," *J. Health Econ.*, Aug. 1983, *2*(2), pp. 167–74.

KEELER, EMMETT B. "What Proportion of Hospital Cost Differences Is Justifiable?" *J. Health Econ.*, Nov. 1990, *9*(3), pp. 359–65.

KEELER, EMMETT B.; CARTER, GRACE M. AND TRUDE, SALLY. "Insurance Aspects of DRG

Outlier Payments," *J. Health Econ.*, Sept. 1988, *7*(3), pp. 193–214.

LAFFONT, JEAN-JACQUES. "Toward a Normative Theory of Incentive Contracts Between Government and Private Firms," *Econ. J.*, 1987 (Conference), *97*, pp. 17–31.

LAFFONT, JEAN-JACQUE AND TIROLE, JEAN. "Using Cost Observation to Regulate Firms," *J. Polit. Econ.*, June 1986, *94*(3, Part 1), pp. 614–41.

———. *A theory of incentives in procurement and regulation*. Cambridge: MIT Press, 1993.

LUFT, HAROLD S. AND MILLER, ROBERT H. "Patient Selection in a Competitive Health Care System," *Health Affairs*, Summer 1988, *7*(3), pp. 97–119.

MA, CHING-TO ALBERT. "Health Care Payment Systems: Cost and Quality Incentives," *J. Economics and Management Strategy*, Spring 1994, *3*(1), pp. 93–112.

MARQUIS, M. SUSAN. "Adverse Selection with a Multiple Choice among Health Insurance Plans: A Simulation Analysis," *J. Health Econ.*, Aug. 1992, *11*(2), pp. 129–51.

MATSAGANIS, MANOS AND GLENNERSTER, HOWARD. "The Threat of 'Cream Skimming' in the Post-Reform NHS," *J. Health Econ.*, Mar. 1994, *13*(1), pp. 31–60.

MCCALL, NELDA AND WAI, H. S. "An Analysis of the Use of Medicare Services by the Continuously Enrolled Aged," *Medical Care*, June 1983, *21*(6), pp. 567–85.

MCCLELLAN, MARK B. "Why Do Medicare Costs Keep Rising? Hospital Reimbursement and the Dynamics of Treatment Intensity." Ph.D. dissertation, Department of Economics, MIT, 1993.

MCGUIRE, THOMAS G. "Price and Membership in a Prepaid Group Practice," *Medical Care*, Feb. 1981, *19*(2), pp. 172–83.

MITCHELL, BRIDGER M. AND PHELPS, CHARLES E. "National Health Insurance: Some Costs and Effects of Mandated Employee Coverage," *J. Polit. Econ.*, May/June 1976, *84*(3), pp. 553–71.

NEIPP, JOACHIM AND ZECKHAUSER, RICHARD J. "Persistence in the Choice of Health Plans," in RICHARD M. SCHEFFLER AND LOUIS F. ROSSITER, eds. 1985, pp. 47–72.

NEWHOUSE, JOSEPH P. "Do Unprofitable Patients Face Access Problems?" *Health Care Financing Rev.*, Winter 1989, *11*(2), pp. 33–42.

———. "Pricing and Imperfections in the Medical Care Marketplace," in *Health economics worldwide*. Eds.: PETER ZWEIFEL AND H. E. FRECH. Amsterdam: Kluwer, 1991.

———. "Patients at Risk: Health Reform and Risk Adjustment," *Health Affairs*, Spring 1994, *13*(1), pp. 132–46.

NEWHOUSE, JOSEPH P. AND BYRNE, DANIEL. "Did Medicare's Prospective Payment System Cause Length of Stay to Fall?" *J. Health Econ.*, Dec. 1988, *7*(4), pp. 413–16.

NEWHOUSE, JOSEPH P. AND THE INSURANCE EXPERIMENT GROUP. *Free for all? Lessons from the RAND health insurance experiment*. Cambridge: Harvard U. Press, 1993.

NEWHOUSE, JOSEPH P. ET AL. "Adjusting Capitation Rates Using Objective Health Measures and Prior Utilization," *Health Care Financing Rev.*, Spring 1989, *10*(3), pp. 41–54.

———. "Risk Adjustment for a Children's Capitation Rate," *Health Care Financing Rev.*, Fall 1993, *15*(1), pp. 39–54.

PAULY, MARK V. "The Economics of Moral Hazard," *Amer. Econ. Rev.*, June 1968, *58*(3), pp. 531–37.

———. *Doctors and their workshops*. Chicago: U. of Chicago Press, 1980.

———. "Taxation, Health Insurance, and Market Failure in Medical Care," *J. Econ. Lit.*, June 1986, *24*(3), pp. 629–75.

POPE, GREGORY C. "Using Hospital-Specific Costs to Improve the Fairness of Prospective Reimbursement," *J. Health Econ.*, Nov. 1990, *9*(3), pp. 237–51.

PRICE, JAMES R. AND MAYS, JAMES W. "Selection and the Competitive Standing of Health Plans in a Multiple-Choice, Multiple-Insurer Market," in RICHARD M. SCHEFFLER AND LOUIS R. ROSSITER, eds. 1985, pp. 127–47.

PRICE, JAMES R.; MAYS, JAMES W. AND TRAPNELL, GORDON R. "Stability in the Federal Employees Health Benefits Program," *J. Health Econ.*, Dec. 1983, *2*(3), pp. 207–23.

ROBINSON, JAMES C. "A Payment Method for Health Insurance Purchasing Cooperatives," *Health Affairs*, Supplement 1993, *12*, pp. 65–75.

ROBINSON, JAMES C. ET AL. "A Method for Risk-Adjusting Employer Contributions to Competing Health Insurance Plans," *Inquiry*, Summer 1991, *28*(2), pp. 107–16.

ROBINSON, JAMES C.: GARDNER, LAURA B. AND LUFT, HAROLD S. "Health Plan Switching in Anticipation of Increased Medical Care Utilization," *Medical Care*, Jan. 1993, *31*(1), pp. 43–51.

ROGERSON, WILLIAM PAUL. "Quality versus Quantity in Military Procurement," *Amer. Econ. Rev.*, Mar. 1990, *80*(1), pp. 83–92.

———. "Choice of Treatment Intensities by a Nonprofit Hospital Under Prospective Pricing," *J. Economics and Management Strategy*, Spring 1994, *3*(1), pp. 7–51.

ROTHSCHILD, MICHAEL AND STIGLITZ, JOSEPH. "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quart. J. Econ.*, Nov. 1976, *90*(4), pp. 629–49.

SCHAUFFLER, HELEN H., HOWLAND, JONATHAN AND COBB, JANET. "Using Chronic Disease Risk Factors to Adjust Medicare Capitation Payments," *Health Care Financing Rev.*, Fall 1992, *14*(1), pp. 79–90.

SCHEFFLER, RICHARD M. AND ROSSITER, LOUIS F., eds. *Advances in health economics and health services research*. Vol. 6. *Biased selection*

in health care markets. Greenwich, CT: JAI Press, 1985.

SELDEN, THOMAS M. "A Model of Capitation," *J. Health Econ.*, 1990, *9*(4), pp. 397–409.

SHLEIFER, ANDREI. "A Theory of Yardstick Competition," *RAND J. Econ.*, Autumn 1985, *16*(3), pp. 319–27.

SIEGEL, CAROLE ET AL. "A Risk-Based Prospective Payment System That Integrates Patient, Hospital, and National Costs," *J. Health Econ.*, May 1992, *11*(1), pp. 1–41.

SUMMERS, LAWRENCE. "Some Simple Economics of Mandated Benefits," *Amer. Econ. Rev.*, May 1989, *79*(2), pp. 177–83.

TOMBERLIN, THOMAS J. "Predicting Accident Frequencies for Drivers Classified by Two Factors," *J. Amer. Statist. Assoc.*, June 1988, *83*(2), pp. 309–21.

U.S. BIPARTISAN COMMISSION ON COMPREHENSIVE HEALTH CARE (THE PEPPER COMMISSION) "A Call for Action." Washington, DC: U.S. GPO, 1990 (S. Prt. 101–114).

VAN DE VEN, WYNAND P. M. M. ET AL. "Risk-Adjusted Capitation: Recent Experiences in the Netherlands," *Health Affairs*, Winter 1994, *13*(5), pp. 120–36.

WEDIG, GERALD J. "Ramsey Pricing and Supply-Side Incentives in Physician Markets," *J. Health Econ.*, Dec. 1993, *12*(4), pp. 365–84.

WEILER, PAUL C. ET AL. *A measure of malpractice*. Cambridge: Harvard U. Press, 1993.

WEINER, JONATHAN P. ET AL. "Development and Application of a Population-Oriented Measure of Ambulatory Care Case-Mix," *Medical Care*, May 1991, *29*(5), pp. 452–72.

WEISBERG, HERBERT I.; TOMBERLIN, THOMAS J. AND CHATTERJEE, SANGIT. "Predicting Insurance Losses Under Cross-Classification," *J. Bus. Econ. Statist.*, Apr. 1984, *2*(2), pp. 170–78.

WELCH, W. PETE. "Restructuring the Federal Employees Health Benefits Program: The Private Sector Option," *Inquiry*, Fall 1989, *26*(3), pp. 321–34.

WILSON, CHARLES A. "A Model of Insurance Markets with Incomplete Information," *J. Econ. Theory*, Dec. 1977, *16*(2), pp. 167–207.

WINSTON, CLIFFORD. "Economic Deregulation: Days of Reckoning for Microeconomists," *J. Econ. Lit.*, Sept. 1993, *31*(3), pp. 1263–89.

ZECKHAUSER, RICHARD J. "Medical Insurance: A Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives," *J. Econ. Theory*, Mar. 1970, *2*(1), pp. 10–26.