Chapter 11

THE ANATOMY OF HEALTH INSURANCE*

DAVID M. CUTLER and RICHARD J. ZECKHAUSER
Harvard University and National Bureau of Economic Research

Contents

*We are grateful to Dan Altman for research assistance, to Jon Gruber, Tom McGuire, Joe Newhouse, and Alexandra Sidorenko for helpful comments, and to the National Institutes on Aging for research support.

Handbook of Health Economics, Volume 1, Edited by A.J. Culyer and J.P. Newhouse

Abstract

This article describes the anatomy of health insurance. It begins by considering the optimal design of health insurance policies. Such policies must make tradeoffs appropriately between risk sharing on the one hand and agency problems such as moral hazard (the incentive of people to seek more care when they are insured) and supplier-induced demand (the incentive of physicians to provide more care when they are well reimbursed) on the other. Optimal coinsurance arrangements make patients pay for care up to the point where the marginal gains from less risk sharing are just offset by the marginal benefits from reduced provision of low valued care. Empirical evidence shows that both moral hazard and demand-inducement are quantitatively important. Coinsurance based on expenditure is a crude control mechanism. Moreover, it places no direct incentives on physicians, who are responsible for most expenditure decisions. To place such incentives on physicians is the goal of supply-side cost containment measures, such as utilization review and capitation. This goal motivates the surge in managed care in the United States, which unites the functions of insurance and provision, and allows for active management of the care that is delivered.

The analysis then turns to the operation of health insurance markets. Economists generally favor choice in health insurance for the same reasons they favor choice in other markets: choice allows people to opt for the plan that is best for them and encourages plans to provide services efficiently. But choice in health insurance is a mixed blessing because of adverse selection – the tendency of the sick to choose more generous insurance than the healthy. When sick and healthy enroll in different plans, plans disproportionately composed of poor risks have to charge more than they would if they insured an average mix of people. The resulting high premiums create two adverse effects: they discourage those who are healthier but would prefer generous care from enrolling in those plans (because the premiums are so high), and they encourage plans to adopt measures that deter the sick from enrolling (to reduce their overall costs). The welfare losses from adverse selection are large in practice. Added to them are further losses from premiums that vary with observable health status. Because insurance is contracted for annually, people are denied a valuable form of intertemporal insurance – the right to buy health coverage at average rates in the future should they get sick today. As the ability to predict future health status increases, the lack of intertemporal insurance will become more problematic.

The article concludes by relating health insurance to the central goal of medical care expenditures – better health. Studies to date are not clear on which approaches to health insurance promote health in the most cost-efficient manner. Resolving this question is the central policy concern in health economics.

Keywords

adverse selection, agency problems, HMOs, indemnity insurance, intertemporal insurance, managed care, moral hazard, pooling equilibrium, separating equilibrium, supplier-induced demand

JEL classification: I10

Insurance plays a central role in the health care arena. More than 80 percent of health care expenditures in the United States are paid for by insurance, either public or private, with an even greater percentage supported in most other developed nations. Insurance thus provides the money that motivates and supports the health care system.

This paper describes the anatomy of health insurance. At the micro level, it details why individuals seek insurance, and the challenges in structuring insurance policies. At the macro level, it explains the role of health insurance in the medical care sector. The medical care triad (Figure 1) depicts that sector in a fundamental fashion. Insurers mediate between individuals[1] and their providers. Often times, the flow of funds is more roundabout: governments or employers nominally pay insurers, but these costs are then passed on to individuals, via increased taxes or lower wages.

The insurer intermediary must design a policy to pay for (and possibly provide) care. This is a treacherous task. Designing a health insurance policy is not nearly so challenging technologically as, say, designing a personal computer system, but it must still overcome some distinct and substantial economic obstacles. The most important of these obstacles are *agency problems*. Insurers cannot get relevant parties to do what efficiency requires. Thus, people with generous insurance spend more on medical care than people with less generous insurance (moral hazard), and providers paid on a fee-for-service (piece-rate) basis may provide more care due to supplier-induced demand than they would if they were not paid per task. In a situation where agency relationships are imperfect, insurance is necessarily second-best. Insurers must trade off the benefits from more generous insurance – primarily the reduction in risk it affords – against the costs of more generous insurance – moral hazard or supplier-induced demand. Throughout this chapter, we highlight central lessons about health insurance, which are then collected in Table 10. This clash between risk sharing and incentives is Lesson 1 about health insurance.

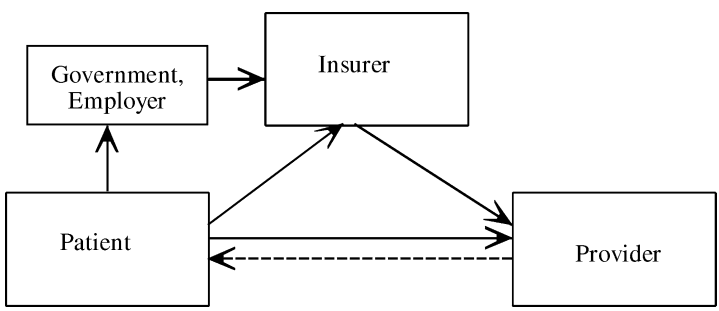


Figure 1. The medical care triad. Solid lines represent money flows; the dashed line represents service flows.

[1] Throughout the paper, to facilitate exposition, we mostly refer to patients or insureds as individuals, although most health insurance is purchased on behalf of families.

Agency problems in health care can be alleviated in two ways. The demand-side approach discourages excessive utilization by making people pay something when they consume medical care. Demand-side rationing is epitomized in the traditional indemnity insurance plan, which prevailed in the United States for a half century. The supply-side approach discourages utilization by monitoring providers carefully, penalizing them if they are profligate, and giving them financial incentives to provide only essential care. Increasingly, supply-side limitations are fostered by integrating insurance and provision. Some HMOs, for example, are both insurers and providers of care. Integration of the insurance and provision functions is unique to medical care, and results from the fundamental difficulties with solely demand-side rationing. The integration of health insurance and provision of medical services is Lesson 2 about health insurance. Sections 3 through 5 of the chapter lay out the issues involved in demand- and supply-side rationing.

We then move from these micro relationships to the broader arena of the market for health insurance. People have preferences for different types of health insurance, and those preferences should be accommodated to the extent possible. In addition, competition in health insurance can encourage production efficiency, driving down overall costs. But competition in health insurance produces results unlike competition in other markets, for a fundamental reason: the costs of providing insurance, as opposed to say computers or food, depend on the characteristics of the buyer. People with a poor medical history will benefit more from and cost more to insure than those with a healthy past. Thus, the sick will sort themselves into more generous plans than will the healthy. This process, called *adverse selection*, can substantially limit the benefits of health plan choice. Individuals will have incentives to choose less generous policies over more generous ones (to pool with the healthy instead of the sick) and insurers will have incentives to reduce the generosity of their benefits (to attract the healthy instead of the sick). Lesson 3 describes the consequences of competition when buyer identity affects costs. Section 6 discusses adverse selection and approaches to deal with it.

The natural tendency of insurers to charge the sick greater premiums than the healthy presents a further challenge to health insurance: lack of coverage against the long-term risk of becoming sick and having higher expected costs in the future. Using the thought experiment of individuals making choices behind the veil of ignorance, they would choose to insure their risk of becoming sicker than average – a multi-year risk – just as individuals in any year wish to insure their medical costs that year. Markets for multi-year insurance do not exist, however, for understandable reasons, and in practice individuals are left without this insurance. The kernel of the problem is that information on risk levels becomes available before insurance contracts are drawn. Lesson 4 is that early information dries up insurance markets. Long-term insurance is taken up in Section 7 of the chapter.

However effectively health insurance controls costs or spreads risks (the focus of most of this chapter), its key goal is to promote health. In Section 8 we examine the relationship between health insurance and health. Variations in insurance generosity have relatively little impact on health outcomes among those with insurance. This finding is

consistent with the idea that insurance generally restricts care offering relatively low value. But the time frame over which these issues has been examined is not large. We know less about the long-run effect of different health insurance arrangements on health than we should. We mark the centrality of health as opposed merely to financial transfers and the lack of clear evidence on the relative benefits of different systems as Lesson 5 about health insurance.

At the outset, it is important to take account of the distinctive role health insurance plays in society. Economists traditionally measure value by willingness to pay, and the value of health insurance, or its byproduct medical care, is calibrated in dollar terms – the same as apples or television sets. In much of the world, however, particularly outside the United States, medical care and medical insurance are treated differently. Medical care is often viewed as a right, for which market-based allocation is not appropriate. For some, the right is absolute; markets should play no role in the allocation of medical services. More moderate positions assign government a special responsibility for medical care, which leads to a government insurance system or set of subsidies. Rights-oriented sentiments show up even in the United States. The United States subsidizes medical insurance directly for poor people and old people, and indirectly for the working-age population (through the exclusion of health insurance from individual taxable income). While some such subsidies may be justified on externality grounds (when people get medical care, they are less likely to spread infectious diseases to others), merit-good arguments, or fiscal externality arguments (when people are healthier, they earn more, pay more in taxes, and receive less in public benefits), we suspect that a right to medical care is the more basic motive.

But the rationale for subsidizing health insurance, as opposed to medical care, is less clear. The government could promote consumption of medical care through direct delivery of services or by subsidizing inputs, without intervening in the medical insurance market. We thus focus primarily on the economic analysis of health insurance, leaving aside normative views about access to basic medical services [Hurley (2000), Wagstaff and van Doorslaer (2000), and Williams and Cookson (2000)]. We come back to the access issue in the last section.

In this essay, we follow common parlance by [primarily] using the terms health care and health insurance, although the terms medical care and medical care insurance might be better descriptors. Health status cannot be insured. The costs of medical care can be, and are, albeit often bearing the label health insurance.

We begin in the first section by discussing the provision of health insurance around the world and in the second with a review of the principles of insurance. We then examine the micro and macro issues in health insurance.

1. Health insurance structures in developed nations

Health insurance is common to all developed countries, but the mechanism for obtaining insurance differs from country to country. In most countries, health insurance is

universal; everyone is entitled to coverage and is required to purchase it.[2] In some nations, such as Canada, the financing is through taxation; people pay an income or payroll tax, and the proceeds are used by the government to purchase or provide health insurance. In other nations, the financing is through private insurance; individuals or their employers contribute to health insurance companies, which then provide insurance for the population. While the payment for any individual may differ in these two systems (a tax-financed system generally imposes relatively more on the rich), the implications for the provision of health insurance are generally slight. Governments in both systems are intimately involved in determining what services are covered, the cost sharing that patients face, and the restrictions imposed on providers.

The specifics of health insurance structures differ significantly across developed nations. Countries such as the UK and Italy finance health insurance through general taxation and (at least historically) provide services publicly.[3] Countries such as Canada and Germany finance insurance publicly but contract for services through private providers.

1.1. Health insurance in the United States

Describing the detailed structures for health insurance in different nations would take an entire volume. We focus our attention primarily on the United States. The United States is distinctive among OECD countries because health insurance is not universal.[4] Table 1 shows the sources of health insurance in the United States. About one-quarter of the United States population is insured through the public sector. The primary public programs are Medicare, which mostly insures the elderly, along with the disabled and people with kidney failure; and Medicaid, which insures younger women and children, the elderly (for services not covered by Medicare such as nursing home care), and the blind and disabled. Other public programs, primarily for veterans and dependents of active-duty military personnel, insure another 1 percent of the population.

Another 60 percent of the population has private health insurance. Most of this insurance is provided by employers; less than 10 percent of the population purchases insurance privately. The predominance of employer-provided insurance results from the favorable tax treatment of that method of payment. Compensation to employees in the form of wages and salaries is taxed through federal and state income taxes, and through the federal Social Security tax. Compensation paid as health insurance, in contrast, goes untaxed. Since marginal tax rates range from 15 to 40 percent for most employees,[5] the

[2] In some countries, such as Germany, temporary workers do not receive health insurance, but they comprise a small part of the population. All citizens are entitled to insurance.

[3] Countries such as the UK have moved to more of a decentralized provision system in recent years. Hospitals have been set up as private trusts, for example, and physicians are no longer salaried.

[4] Since 1996, health insurance coverage has been required in Switzerland, but before then it was subsidized so heavily that essentially everyone purchased it.

[5] Income tax rates can range as high as 40 percent, but the income level at which these rates are reached are past the cap on earnings subject to the payroll tax.

Table 1
Sources of health insurance coverage for the United States population

Source	Groups insured	Share of total population (%)	Share of total payments (%)
Public			
Medicare	Elderly; disabled; end-stage renal disease	13	22
Medicaid	Elderly; blind and disabled; poor women and children	10	15
Other*	Military personnel and their dependents	1	8
Private			
Employer sponsored	Workers and dependents	56	53
Nongroup	Families	6	
Uninsured		16	2

* Other public spending includes non-insurance costs such as public hospitals, the Veterans Administration, etc.

Source: Authors' calculations based on data from Department of Health and Human Services, National Health Accounts (medical spending), and from Employee Benefit Research Institute (insurance coverage).

subsidy to employer-provided insurance, as opposed to individually-purchased insurance, is substantial. The subsidy to employer-provided health insurance generally does not extend, however, to out-of-pocket payments made by employees. As a result, there are incentives to have generous insurance, paid for by employers, with few individual copayments. We return to the effects of this subsidy structure below.

The remaining 16 percent of the United States population is uninsured. The implications of being uninsured are a subject of vigorous debate [Weissman and Epstein (1994)]. Some of the uninsured (perhaps 4 percent) are eligible for public insurance (particularly Medicaid) but have chosen not to take up that insurance. Presumably, if these people become sick they will enroll in Medicaid.[6] Others will receive "uncompensated" care if they become sick – they will get emergency care if they need it, but they will not pay for it. The costs of uncompensated care then get shifted to people with insurance, for whom payments made exceed the cost of services provided. In this sense, the United States has a form of universal insurance coverage for catastrophic care, although the patchwork nature of that coverage is undoubtedly suboptimal. It also limits primary and preventive care for those without health insurance.

The last column of Table 1 shows the share of total payments that each group makes. As in any insurance policy, people may use more or less of the service than they pay

[6] Since it is difficult to deny treatment, providers have a strong interest to enroll eligible people in Medicaid, so that they can receive some payment for them.

for. This is particularly true for the uninsured, whose out-of-pocket payments are much lower than the cost of services they receive. The table reports the share of total payments made by each group; the share of services that is used by each group will be somewhat different. Because people insured through the public sector are older and sicker than people insured privately, and because some of the costs of the uninsured are passed on to the public sector, the public sector accounts for much more of medical spending than its demographic share of insurance coverage. Close to half of medical spending in the United States is paid for publicly. While this amount is extremely high relative to most goods and services in society, it is low by international standards for medical care. In OECD nations, governments generally pay for 75 to 90 percent of medical care.

Whether run publicly or privately, health insurance encounters fundamental problems that any insurer must face. Adverse selection, though diminished for government since some of its programs are so heavily subsidized that the vast majority choose to participate, still exists, and moral hazard affects governments no less than private insurers. Thus, when we discuss the optimal design of health insurance policies, we do not distinguish between public or private insurers. We return to public versus private insurance issues in the conclusion.

2. The principles of insurance

In this section and the next three, we discuss the optimal design of health insurance policies. Our perspective is that of an insurer – public or private – wanting to optimally insure its enrollees against the costs of treating adverse health outcomes.

The value of health insurance is rooted in the unpredictability of medical spending. While individuals know something about their need for medical services, the exact amount they will spend on medical care is to a significant degree uncertain. Medical spending is extremely variable. Table 2 shows the distribution of medical spending in the United States in 1987 [Berk and Monheit (1992)]. The top 1 percent of medical care users consume an average of nearly $50,000 each in a year (in 1987 dollars), and

Table 2
Distribution of medical spending, 1987

Share of distribution	Cumulative share of spending (%)
Top 1 percent	30
Top 5 percent	58
Top 10 percent	72
Top 50 percent	98
Total population	100

Source: Berk and Monheit (1992).

account for 30 percent of medical spending. The top 10 percent of users account for nearly three-quarters of total medical spending. The shorter the time period, of course, the greater is the percentage disparity in medical spending among individuals. But even looking over several years, the skewness of medical spending is substantial [Roos et al. (1989), Eichner, McClellan, and Wise (1998)]. In such a situation, insurance can significantly spread risks.

Risk-averse individuals will want to guard against the potential of requiring a substantial amount of medical care. One way to do this is to wait, borrow money for treatment should they get sick, and then repay the money when well. But borrowing when debilitated is difficult, since the individual may not live long enough or be healthy enough to repay the loan. The borrowing process, moreover, may also take more time than the sick individual has available. A reasonable alternative might be for individuals to save money when they are healthy to pay for medical care should they get sick. But some sicknesses are significantly more expensive than others. The substantial expenses of very severe illness make saving prior to illness impractical as a protective measure. All of us would have to significantly curtail consumption to save up for expenses that would be borne by only a few. The natural solution is to insure against the possibility of medical illness by pooling risks with others in the population. Annual consumption would be reduced only by the premium, the average cost of care.

Risks to health have always been with us, but health insurance is a relatively new phenomenon, only becoming economically significant in the postwar era. Fire and life insurance were well developed by the end of the 19th century, and marine insurance was already being written in the 12th century. There was little role for health insurance in earlier eras, however, since expensive medical treatments could accomplish little for health. Insurers also feared they could not control individual use of medical services if the services were insured. Once effective hospital care – an extremely expensive commodity – became possible, significant health insurance became desirable and inevitable.

2.1. Insurance with fixed spending

The simplest insurance situation is one where sickness entails a fixed cost and insurance is priced at its actuarial cost. Imagine a situation where initially identical individuals are either healthy or sick in a period of one year. There is one disease. People are healthy with probability $1 - p$, in which case they require no medical care. People get sick with probability p. Let $d = 0$ or $d = 1$ indicate whether absent medical care the person is healthy or sick. Treatment of a person who is sick requires medical spending of m. The after-expenditure health of a sick person is $h = H[d, m]$. To simplify exposition, we assume that medical spending restores a person to perfect health, so that $H[1, m] = H[0, 0]$.

Before proceeding, we alert the reader to our use of mathematics. We use mathematics to derive statements precisely. We also endeavor to explain all of our results intuitively. Thus, readers who wish to skip the mathematical portions of the chapter can still follow the central arguments.

Individuals receive utility, u, which depends on their consumption, x, and their after-treatment health, h. Thus we have $u = U(x, h)$. Assume, for simplicity, that people have exogenous income endowments, y; and that they can neither borrow or lend. Thus, an individual's consumption is what is left over after paying medical expenditures, or if insured, his insurance premium, π. Thus, for uninsured people, $x = y$ when healthy and $x = y - m$ when sick. For insured people, $x = y - \pi$ whether healthy or sick. We use the subscripts I and N to indicate whether the individual is insured or not insured.

Let $U(x) \equiv U(x, H[0, 0])$; i.e., it is the reduced form utility function for consumption given perfect health. In the absence of insurance, an individual's expected utility is given by:

$$\begin{aligned} V_{\mathrm{N}} &= (1-p)U\big(y, H[0,0]\big) + pU\big(y-m, H[1,m]\big), \\ &= (1-p)U(y) + pU(y-m), \end{aligned} \tag{1}$$

where the second equality follows from the assumption that medical care restores the person to perfect health.[7] We assume that U has the standard property that utility is increasing in consumption albeit at a declining rate: $U' > 0$ and $U'' < 0$. We further assume that medical expenditures are worthwhile even if the individual is not insured.

Suppose the individual purchases insurance against the risk of being sick. For an insurance company to break even, the fair insurance premium would have to be $\pi = pm$. The insurance company collects the premium each year and pays out m when the individual is sick. If an individual chooses this policy, his utility would always be:

$$V_{\mathrm{I}} = U(y - \pi). \tag{2}$$

Using a Taylor series expansion of Equation (1),[8] we can approximate that equation as:

$$V_{\mathrm{N}} \approx U(y-\pi) + U'\big(U''/2U'\big)\pi(m-\pi). \tag{3}$$

Therefore,

$$\text{Value of Insurance} = (V_{\mathrm{I}} - V_{\mathrm{N}})/U' \approx (1/2)\big(-U''/U'\big)\pi(m-\pi). \tag{4}$$

[7] Assuming that medical expenditure is worthwhile, this analysis actually requires a less stringent condition. The same equation would apply if restored health imposed a fixed utility cost, k, relative to initial perfect health, so that $U(c, H[0,0]) = U(c, H[1,m]) + k$ for all c.

[8] The Taylor series is taken about the level of income net of insurance premiums. From Equation (1), $V_{\mathrm{N}} \approx (1-p)[U(y-\pi) + U'\pi + (1/2)U''\pi^2] + p[U(y-\pi) - U'(m-\pi) + (1/2)U''(m-\pi)^2]$. Collecting terms, this simplifies to $V_{\mathrm{N}} \approx U(y-\pi) + U'\{(1-p)\pi - p(m-\pi)\} + (1/2)U''\{(1-p)\pi^2 + p(m-\pi)^2\}$. The term $(1-p)\pi - p(m-\pi)$ is zero. The term $(1-p)\pi^2 + p(m-\pi)^2$ can be expanded as $(1-p)\pi^2 + pm^2 - 2pm\pi + p\pi^2$. Since $pm = \pi$, this simplifies to $pm^2 - \pi^2 = \pi(m-\pi)$.

The left hand side of Equation (4) is the difference in utility from being uninsured relative to being insured, scaled by marginal utility to give a dollar value for removing risk. The right hand side is the benefit of risk removal. Here, $(-U''/U')$ is the *coefficient of absolute risk aversion*; it is the degree to which uncertainty about marginal utility makes a person worse off. Because $U'' < 0$ and $U' > 0$, this term is positive. The term $\pi(m - \pi)$ represents the extent to which after-medical expenditure income varies because the person does not have insurance. It too is positive. The product of terms on the right hand side of Equation (4), therefore, is necessarily positive, implying that fair insurance is preferred to being uninsured. The dollar value of risk spreading increases with risk aversion and with the variability of medical spending.

The intuition supporting this result is that risk averse individuals would like to smooth the marginal utility of income – to transfer income from states of the world where their marginal utility is low to states of the world when their marginal utility is high. In the absence of insurance, a person's marginal utility of income when healthy is $U'(y)$ and when sick is $U'(y-m)$. Since marginal utility falls as income increases, marginal utility is lower when healthy than when sick. Transferring income from healthy states to sick states until marginal utility is equalized maximizes total utility, assuming fair insurance. Health insurance carries out this transfer, charging premiums up front and reimbursing expenditures later.[9]

There is a diagrammatic way to make the same point; it is shown in Figure 2. We think of the two states of the world – being sick and being healthy – as if they were two goods. Individuals would like more consumption in each state. In the absence of any probability of being sick, people would be able to consume y in each state. Because of required medical spending, however, people can only consume $y - m$ when sick. This is shown as point E in the figure.

[9] The situation is more complex when medical spending fails to restore the person to perfect health, and the marginal utility of income is affected by health status. Suppose that when sick a person still needs medical spending of m, but that his after-expenditure health remains below what it would be had he never got sick; i.e., that $H[1, m] < H[0, 0]$. Expected utility for people without insurance is given by $V_N = (1 - p)U(y, H[0, 0]) + pU(y - m, H[1, m])$, and the marginal utilities of income are U_x $(y, H[0, 0])$ when healthy and $U_x(y - m, H[1, m])$ when sick, where the subscripts indicate partial derivatives. Because the marginal utility of income may be affected by health and health varies across sickness states, it is not clear how much insurance the person will want. If people attach little value to money when sick – for example, if there are few pleasurable activities they can engage in – they may not want any health insurance at all. Alternatively, if the value of money when sick is particularly high, say because aides are needed to carry out the activities of daily life, people may want more than full insurance against medical expenditures.

This example highlights the difference between *medical care insurance* and what, if we used a strict interpretation, would be labeled *health insurance*. Health insurance transfers money across people – generally from the healthy to the sick. The money can be used to purchase medical services the individual otherwise could not afford, or to allow the individual to purchase more of other goods and services after medical care has been paid for. But health insurance cannot guarantee that an individual's health will be unaffected by outside factors. Insuring one's health is technologically infeasible.

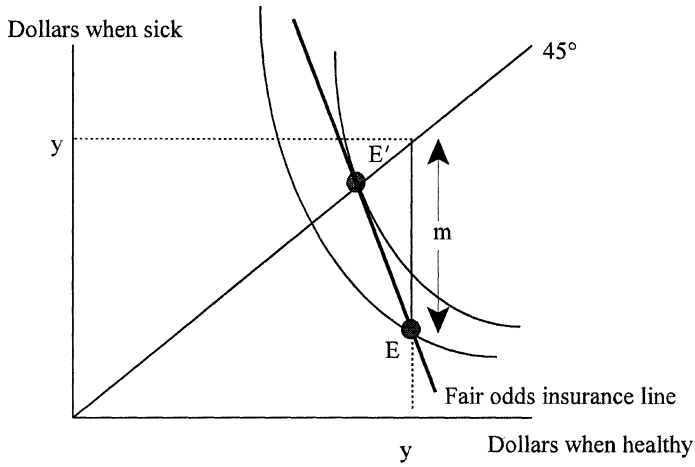


Figure 2. The welfare gains from health insurance.

The fair odds insurance line is the individual's implicit budget constraint. It is drawn for the case where $p = 0.2$. The slope of the line is $-1/p$, or -5.[10] The indifference curve for consumption is also steeply sloped, recognizing that the sick state is unlikely to arise. Thus, people are not willing to give up much consumption when healthy to get consumption when sick. A person can trade consumption when sick for consumption when healthy, at a rate given by the insurance premium. People will choose to purchase some insurance. If insurance is priced actuarially fairly, individuals will choose to be fully insured – they will have the same consumption when sick as when healthy. This optimum is shown as point E′ in the figure. People are better off at E′ than they are at E; they have moved to a higher indifference curve.

In our simplified world, the optimal insurance policy is an *indemnity* policy – it pays a fixed amount of money for a particular condition when the individual is sick. The amount paid equals the cost of the appropriate treatment for the person's disease; if there is more than one disease, the payments vary. Since each disease requires a fixed amount of care – there is no more nor less that a person can consume – there are no wasted resources in the policy; the indemnity insurance plan is efficient. Beyond its efficiency properties, the indemnity policy is the simplest health insurance policy. In effect, it operates as a contingent claims market; people get paid a specified amount depending on which contingency occurs [Zeckhauser (1970)].

Health insurance started off as a quasi-indemnity policy – in most cases paying a fixed amount per day in the hospital. The first Blue Cross policies, for example, were

[10] A fair insurance policy that charges \$1 each year and pays an amount k when sick is defined by: $pk + 1 = 0$. Thus, $k = -1/p$. Some authors assume the insurance payment is made only when the person is healthy, in which case the fair odds policy is defined by: $pk' + (1 - p) = 0$, or $k' = -(1 - p)/p$.

developed just before and during the Great Depression. These policies, run by hospitals, guaranteed a certain number of hospital days per year (for example, 21 days) for an annual premium (for example, $5 to $10 in the early 1930s). After World War II, life insurance companies entered the health insurance market, driven by the profits of Blue Cross policies and the expanding demand for health insurance resulting from its favorable tax treatment. These nascent health insurers offered indemnity policies as well, limiting their potential losses by fixing the maximum amount they would pay per hospital day.

3. Moral hazard and principal-agent problems

Health insurance must address several problems beyond risk spreading. We now turn to some of these challenges.

3.1. Moral hazard

Moral hazard refers to the likely malfeasance of an individual making purchases that are partly or fully paid for by others [Arrow (1965), Pauly (1968, 1974), Zeckhauser (1970), Spence and Zeckhauser (1971), Kotowitz (1987)].[11] He will overspend; i.e., he will use more services than he would were he paying for the medical care himself. Since insurance is an arrangement where others pay for the lion's share of one's losses, it creates a moral hazard to use additional medical resources. The designation moral hazard, a disquieting term, frequently connotes some moral failure of individuals, but this is not meant to be so. Indeed, Kenneth Arrow (1985) employs the less judgmental and more informative term "hidden action" for moral hazard.

Moral hazard is a concern because it conflicts with risk-spreading goals. Insurance is valuable because it allows people to transfer income from when they need it less to when they need it more. But this transfer is not perfect because people increase their consumption of medical care when it is subsidized. This creates an inherent second-best problem in designing insurance policies: insurers must trade off the benefits from spreading more risk against the cost of increased moral hazard. We formalize this Lesson 1 about health insurance:

> *Lesson* 1: *Risk spreading versus incentives*. Health insurance involves a fundamental tradeoff between risk spreading and appropriate incentives. Increasing the generosity of insurance spreads risk more broadly but also leads to increased losses because individuals choose more care (moral hazard) and providers supply more care (principal-agent problems).

[11] The theory of moral hazard, if not the words, goes back at least to Adam Smith: "The directors of such companies, however, being the managers rather of other peoples' money than of their own, it cannot well be expected, that they should watch over it with the same anxious vigilance with which the partners in a private copartnery frequently watch over their own... Negligence and profusion, therefore, must always prevail, more or less, in the management of the affairs of such a company" [Smith (1776, p. 700)].

Moral hazard, or hidden action, emerges in one form in the risks that individuals choose to take. People may take worse care of themselves when they have insurance than if they do not. If their actions were readily observed, the insurance company would merely not pay off were they reckless or negligent. But individual actions are difficult to observe; they are hidden. The extent of moral hazard in terms of actions that affect health may not be large for health insurance in most instances, since the uncompensated loss of health itself is so consequential.[12] Thus, it would be surprising if people smoked because they knew health insurance would cover the costs of lung cancer.

Hidden action also arises because individuals may get treatments they would not pay for themselves. Though the action itself (seeking medical care) is not hidden, the motivation behind it is.[13]

Optimal insurance plans would pay for treatment only if the individual would have chosen the same treatment had he borne the full bill. The thought experiment here is whether the person would pay for the medical expenditure in expectation, before he knew his condition. For example, suppose that a person has income of $25,000, and faces a 1 percent probability he will have a serious illness. If he could commit in advance, he would agree to receive $50,000 of medical care when sick in exchange for a $500 premium. If fully insured, however, the individual will choose to consume $60,000 of care. The moral hazard in this example is $10,000 – the additional spending beyond the optimal amount of care he would contract for in advance of being sick.

In the terminology of demand theory, moral hazard is the *substitution effect* of people spending more on medical care when its price is low, not the *income effect* of people spending more on medical care because of insurance, by efficiently transferring resources from the healthy state to the sick state, makes them richer when sicker [De Meza (1983)]. In the example considered, say the individual would have spent half his income, $12,500, on medical care in the absence of insurance. Insurance thus raises medical spending by $47,500, but only a fraction of this increase is due to moral hazard.

If some fixed m were the known optimal medical expenditure for any sick person, insurance plans would experience no moral hazard. They could simply pay m in medical expenditures to or for those who are sick. Moral hazard arises because medical needs are not fully monitorable, and different people with the same condition have different optimal expenditures, at least as best the insurance company can determine. Suppose that the optimal medical expenditure for treating a particular condition is m_i, which varies across people, indexed by i. The insurance company requires the individual to pay a coinsurance amount $c(m)$ for medical care received. The rest of the care, $m - c(m)$, is paid by the insurer. In effect, the insurer takes the individual's medical expenditure

[12] This does not mean that people will not smoke or faithfully take their medications. But there is no moral hazard if their actions would be the same if they had no health insurance, i.e., if these health-harming behaviors are inelastic with respect to cost sharing.

[13] Moral hazard also results from patients making less effort to search for low-cost providers. For example, when patients pay but one-fifth of the cost of their drugs, they will have weak incentives to switch to generic brands or stray beyond the local pharmacy.

to be a signal of his true medical needs; the coinsurance payment creates the necessary costs to have signaling operate.

Two polar extremes for the form of $c(m)$ are commonly found. The first is the indemnity policy discussed above: the insurer pays a fixed amount, call it m^*, and the individual pays $c(m) = m - m^*$. The second is full insurance: the insurer pays the full costs of medical care, regardless of its cost, and the individual pays nothing (i.e., $c(m) = 0$). The full insurance policy removes all risk from the insured, but engenders greater moral hazard.

To understand the optimal insurance policy, consider a case where an indemnity policy is not optimal. Suppose that rather than being healthy or sick, the individual has a range of potential illness severities, s, with s distributed with density function $f(s)$. Health is given as before by $h = H[s, m]$. The patient's s will determine the optimal treatment. The insurer cannot observe s, however. Thus, making a fixed indemnity payment to anyone sick is not optimal. The *ex ante* utility function for the insured consumer is:

$$V_{\mathrm{I}} = \int U\big(y - \pi - c(m(s)),\, H[s, m(s)]\big) f(s)\, \mathrm{d}s, \tag{5}$$

where $m(s)$ tells how much medical care an individual with condition s chooses to receive.

We consider first the optimal policy – the amount of medical services the person would like to contract for if he could write a perfect state-contingent contract and thereby eliminate moral hazard. When s is observable, the coinsurance rate depends only on s, hence can be written as $c(s)$. The individual will choose $m^*(s)$ maximum feasible utility:

$$\mathrm{Max}_{m(s)} \int U\big(y - \Pi - c(s),\, H[s, m]\big) f(s)\, \mathrm{d}s, \tag{6}$$

where $\Pi = \int (m(s) - c(s)) f(s)\, \mathrm{d}s$. The solution to this problem sets

$$H_m U_H = E[U_x], \tag{7}$$

where the subscripts denote partial derivatives and $x = y - \Pi - c(s)$. The left-hand side represents the gain in utility from spending another dollar on medical care; it is the product of the effect of medical care on health and the effect of health on utility. The right hand side is a weighted average expectation of the marginal utility of consumption in different illness states, namely:

$$E[U_x] = \int U_x\big(y - \Pi - c(s),\, H[s, m]\big) f(s)\, \mathrm{d}s. \tag{8}$$

Equation (7) says that with the optimal first-best policy, the expected marginal utility gained from an additional dollar of medical care in each state of the world equals the utility cost of a dollar.[14]

In the case where the marginal utility of income does not depend on the health state,[15] imposing a coinsurance payment in any health state, i.e. a variable $c(s)$, increases the variability of income and thus reduces expected utility. The optimal policy for this commonly studied case in thus no coinsurance, and a payment $m^*(s)$ that fully reimburses optimal spending in each state.[16]

Now consider a situation where severity of illness is not monitorable, hence the optimal policy just discussed cannot be implemented. At the time the consumer is seeking medical care, he alone knows his severity. We assume the consumer treats the insurance premium as fixed – nothing he does will raise or lower his insurance premium that year. Further, we assume for now that individuals are not penalized in future years for additional medical spending this year, because expected future changes in costs are spread equally over everyone in the group. The cost to the consumer of another dollar of medical expenditure will be $c'(m)$.[17] The sick consumer will therefore choose medical care utilization to maximize utility when sick. Thus, he will choose $m^{\#}(s)$ as the m which maximizes utility given knowledge of s:

$$\text{Max}_{m(s)} U\big(y - \Pi - c(m), H[s, m]\big) \quad \text{for each } s. \tag{9}$$

The solution to this problem will depend on the specific s the individual has realized, and is given by the first order condition:

$$H_m U_h = c'(m) U_{zx} \quad \text{for each } s. \tag{10}$$

The left-hand side once again represents the gain in utility from spending another dollar on medical care. The right-hand side is the utility cost to the individual from spending

[14] This assumes that these functions are well behaved, hence that local optima are global optima. Some medical expenditures may offer increasing returns over a relevant range. For example, it may cost $200,000 to do a heart transplant, with $100,000 accomplishing much less than half as much. Efficiency then requires the insurance program spend at least to the minimum average cost of benefits point, or not at all.

[15] This case would arise if utility is additively separable between income and health.

[16] If utility does depend on the health state, for example, if a disabled person needs more non-medical services, then optimal coincurance will actually pay the individual when disabled.

[17] The structure of the insurance plan may present the insured with a range of decreasing marginal cost. Say a plan has a deductible of $600 with a copayment of 20% beyond that point, a common structure. The insured can receive $600 of benefits for $600, but $1200 of benefits for $720. Say the individual solves, and finds a $400 expenditure is locally optimal. He must also look globally to the optimal expenditure beyond $600, which may be superior. Recognizing that using up a deductible gets one to a range of lower costs, gives the insured an interesting dynamic optimization program where there are two benefits from spending below the deductible: (1) the health care itself, and (2) the increased potential for getting to the low-cost range [Keeler, Newhouse, and Phelps (1977)].

that dollar; it is the product of the out-of-pocket cost of medical care and the utility loss from losing that dollar for consumption.

Comparing Equations (7) and (10) shows the loss due to moral hazard. When $c'(m) < 1$, as it will be when marginal spending is in any way insured, people will overconsume medical care when sick and thus pay more for health insurance than is optimal.[18]

3.1.1. Evidence on the price elasticity of medical care demand

How does an individual's demand for medical care respond to his required out-of-pocket expenses? Economists used to differ on this question. Table 3 details estimates of the elasticity of demand for medical care.[19] A substantial literature in the 1970s estimated the elasticity of demand for medical care using cross-sectional data, or cross-sectional time series data. Pre-eminent among these papers are Feldstein (1971), Phelps and Newhouse (1972b), Rosett and Huang (1973), and Newhouse and Phelps (1976). Feldstein (1971) was the first statistically robust estimate of price elasticities using time-series micro data, in this case on hospitals. Feldstein identified the effect of coinsurance rates on demand using state-variation in insurance coverage and generosity, estimating a demand elasticity of about -0.5. The subsequent papers use patient-level data and more sophisticated study designs. The elasticities that emerged from these papers ranged from as low as -0.14 [Phelps and Newhouse (1972b)] to as high as -1.5 [Rosett and Huang (1973)]. The implication of this range of elasticity estimates was that moral hazard was likely a significant force.

This estimation literature suffered from two major difficulties, however. First, the generosity of health insurance at either the state or the individual level might be endogenous. Generous insurance might boost utilization of medical services, as posited; or alternately, areas where people desire or need more medical care may also be areas where people demand more health insurance. One cannot separate these two effects statistically without an instrument for the rate of insurance coverage in an area, but such instruments were not easy to find. Second, the studies typically failed to distinguish average and marginal coinsurance rates. Usually for data reasons, most of these studies related medical spending to the *average* coinsurance rate in an area. But theory predicts that medical spending should relate to the *marginal* coinsurance rate. Because insurance policies are non-linear, average and marginal prices may differ substantially.[20] As a result of these problems, as late as the 1970s many critics still believed that medical care was determined by "needs" and no other economic factors, i.e., that demand was totally

[18] This can be derived by taking expectations of both sides of Equation (10) and comparing to Equation (7). There is also a risk-bearing loss when severity, is not monitorable, as reflected by the term U_z in (10) as opposed to $E(u_x)$ in (7).

[19] Zweifel and Manning (2000) discuss the elasticity of demand for medical care in more detail.

[20] Of course, if individuals are appropriately forward looking, it is the *expected* marginal coinsurance rate at the end of the year that should affect behavior, rather than the ostensible marginal coinsurance rate at the time services are used.

Table 3
Estimates of the elasticity of demand for medical care

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Feldstein, P.J. (1964)	1953, 1958 Health Information Foundation and NORC surveys	general care	cross-section estimates of physician visits	−0.19 (physician visits)		
Feldstein, M.S. (1970)	BLS survey; NCHS 1963–1964 survey; physician interviews	aggregated physician service data	time-series regression	1.67 (physician services)		
Rosenthal (1970)	1962 sample of New England hospitals	68 of 218 general, short-term hospitals	univariate estimates for short-term care categories	0.19 to −0.70		
Feldstein, M.S. (1971)	AHA survey of hospitals, 1958–1967, NCHS 1963–1964 survey	all hospitals, aggregated by state	time-series regression	−0.49 for total bed days	−0.63 for visits to hospital	
Davis and Russell (1972)	1970 guide issue of "Hospitals"	aggregated hospital outpatient care; 48 states' not-for-profit hospitals	cross-sectional estimates	−0.32		
Fuchs and Kramer (1972)	1966 Internal Revenue Service tabulations	physician services, aggregated into 33 states	TSLS: IVs are number of medical schools, ratio of premiums to benefits, and union members per 100 population	−0.10 to −0.36		
Phelps and Newhouse (1972a, 1972b)	Palo Alto Group Health Plan, 1966–1968	physician and outpatient ancillary services	natural experiment: introduction of coinsurance	−0.14* OLS, −0.118 Tobit (physician visits)		

continued on next page

Table 3, *continued*

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Scitovsky and Snyder (1972)	Palo Alto Group Health Plan, 1966–1968	physician and outpatient ancillary services	natural experiment: introduction of coinsurance	−0.060* (ancillary)	−0.14* (physician visits)	
Phelps (1973)	verified data from 1963 CHAS (University of Chicago) survey	hospitalization and physicians' services	cross-sectional Tobit estimates	not significantly different from zero		
Rosett and Huang (1973)	1960 Survey of Consumer Expenditure	hospitalization and physicians' services	cross-sectional Tobit estimates	−0.35 to −1.5		
Beck (1974)	random sample of poor population of Saskatchewan	physicians' services	natural experiment; introduction of co-payments	−0.065*		
Newhouse and Phelps (1974)	1963 CHAS survey	employeds' hospital stays within coverage	cross-sectional OLS (TSLS estimates insignificant)	−0.10 (length of stay)	−0.06 (physician visits)	
Phelps and Newhouse (1974)	insurance plans in US, Canada, and UK	general care, dental care, and prescriptions	arc elasticities across coinsurance ranges	−0.10		
Newhouse and Phelps (1976)	1963 CHAS survey (larger sample than in previous work)	employeds and non-employeds	cross-sectional OLS (TSLS estimates insignificant)	−0.24 (hospital), −0.42 (physician)		
Scitovsky and McCall (1977)	Palo Alto Group Health Plan, 1968–1972	physician, outpatient ancillary services	natural experiment: coinsurance increases	−2.56* (ancillary)	−0.29* (physician visits)	
Colle and Grossman (1978)	1971 NORC/CHAS health survey	pediatric care	cross-sectional estimates	−0.11	−0.039	

continued on next page

Table 3, *continued*

Paper	Data	Restrictions	Estimation method	Total price elasticity	Visits price elasticity	Quality price elasticity
Goldman and Grossman (1978)	1965–1966 Mindlin–Densen longitudinal study	pediatric care	hedonic model		−0.060 (compensated −0.032)	−0.088 (compensated −0.085)
McAvinchey and Yannopoulos (1993)	waiting lists from UK's National Health Service	acute hospital care	dynamic intertemporal model	−1.2		
Newhouse et al. (1993)	RAND Health Insurance Experiment	general care	randomized experiment	−0.17 to −0.31 (hospital), −0.17 to −0.22 (outpatient)		
Bhattacharya et al. (1996)	1990 Japanese Ministry of Health and Welfare survey	outpatient visits	Cox proportional hazards model	−0.22		
Cherkin et al. (1989)	Group Health Cooperative of Puget Sound	non-Medicare HMO patients	natural experiment: introduction of copayments	−0.035* (all visits), −0.15* to −0.075* (preventive)		
Eichner (1998)	1990–1992 insurance claims from employees and dependents of a Fortune 500 firm	employees aged 25 to 55	one-and two-stage Tobit regressions of out-of-pocket costs	−0.32		
SUMMARY				−0.20	−0.05 to −0.15	

* Elasticities computed according to appendix of Phelps and Newhouse (1972b).

inelastic, although others believed that the demand elasticity was substantial – perhaps −0.5 or more.

To address these problems, the United States government funded a social insurance experiment, designed to estimate the demand elasticity for medical care. The Rand Health Insurance Experiment [Newhouse et al. (1993), Zweifel and Manning (2000)] randomized nearly 6,000 people in 6 areas to different insurance plans over a 3- to 5-year period in the early 1970s. The insurance plans varied in contractual levels of cost sharing. Elasticity estimates were formed by comparing utilization in the different plans. The Rand Experiment found an overall medical care price elasticity of about −0.2. This elasticity is statistically significantly different from zero, but noticeably smaller than the prior literature suggested. Sound methodology, supported by generous funding, carried the day. The demand elasticities in the Rand Experiment have become the standard in the literature, and essentially all economists accept that traditional health insurance leads to moderate moral hazard in demand. The Rand estimates are also commonly used by actuaries in the design of actual insurance policies.

3.1.2. Coinsurance in practice

The indemnity policy, which characterized health insurance at its inception, became outdated over time. With increased medical technology, the range of optimal spending within a given condition became great. Indemnity policies left individuals bearing too much risk. As a result, insurance structures moved from indemnity payments to a service benefit policy – a policy that covers all medical expenses, with some cost sharing. Service benefit policies grew steadily in importance in the post-war period, reaching their height in the early 1980s.

Service benefit policies use three cost-sharing features, sometimes in concert: the *deductible*, the *coinsurance rate*, and the *stop loss* amount. Figure 3 and Table 4 show how these cost sharing features operate. The deductible is the amount that an individual must pay before the insurance company pays anything. The deductible is usually set annually; the typical deductible in 1991 was about $200 for an individual and $500 for a family. Consumers pay the full price for care consumed under the deductible. The coinsurance rate is the percentage of the total bill above the deductible that a patient pays. Nearly all indemnity plans had a coinsurance rate of 20 percent. The coinsurance is paid until the patient reaches the stop loss – the maximum out-of-pocket payment by the person in a year. A typical stop loss in an indemnity policy was about $1,000 to $1,500 in a year.

In addition to these features, many policies impose further cost sharing through caps on various types of expenditures. For example, policies may permit 8 mental health visits per year, or have a $1 million lifetime limit on overall medical expenditures. Such provisions discourage use, and may deter high cost users from selecting the insurance plan, and providers from turning expensive cases into subsidized meal tickets. Table 4 details the frequencies with which various policy features were found in insurance policies in 1991.

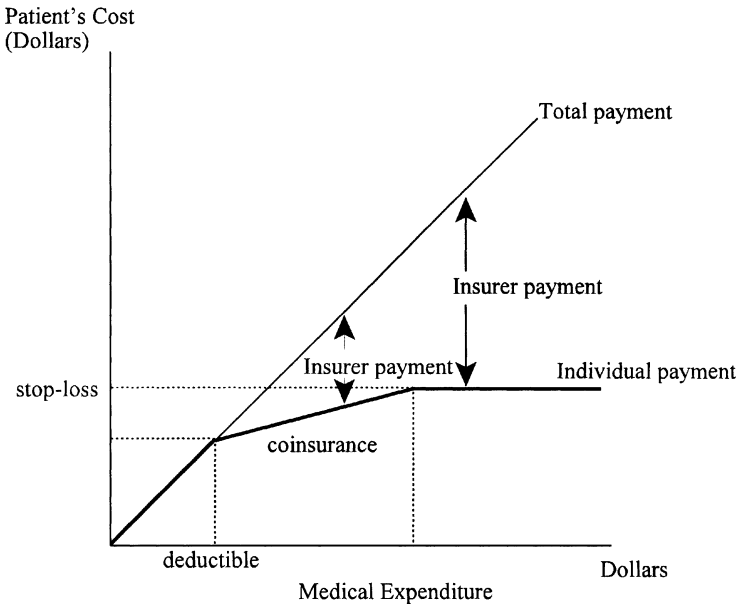


Figure 3. Cost sharing under indemnity insurance.

Table 4
Risk-sharing features of indemnity insurance policies, 1991

Characteristic	Average/percent
Deductible	
Individual	$205
Family	$475
*Coinsurance rate**	
<20 percent	13%
20 percent	78%
>20 percent	4%
Stop loss	
⩽$500	21%
$501–$1,000	30%
$1,001–$2,000	32%
>$2,000	17%
Maximum lifetime benefit – individual	
⩽$250,000	9%
$250,001–$999,999	6%
⩾$1,000,000	85%

Source: HIAA Employer Survey, 1991.
* Remaining responses are "rate varies" and "other".

Somewhat misleadingly, the service benefit policy is frequently called an "indemnity insurance plan" by economists, with the system that developed to provide this policy termed the "indemnity insurance system". In fact, true indemnity health insurance policies (a fixed payment per disease) had existed but were largely replaced by the service benefit policy. For consistency with other literature, we follow this nomenclature despite its inaccuracies. This nomenclature is particularly unfortunate since recently insurance has been moving back towards the indemnity model, frequently with the risk of above-average spending being placed on the provider rather than the patient. We discuss regimes of provider responsibility in Section 4.

3.1.3. Optimal insurance given moral hazard

Knowledge of the utility function and the parameter values that determine medical spending elasticities can be combined to design the optimal insurance policy – the actuarially fair policy that maximizes expected utility subject to the constraint that individuals will act in a self-interested fashion, i.e., that moral hazard will operate. Such a policy is inherently second-best; in calibrating its level of generosity, it balances the utility benefits of greater risk-sharing across people against the moral hazard costs incurred. The insurer's challenge is to define the function of risk sharing by insureds, the $c(m)$ function, that maximizes expected utility.

To analyze the optimal policy, we assume patients differ in the severity of their illness.[21] The insurer will seek to find the $c^*(m^{\#})$ function that produces the maximum possible expected utility with:

$$E[U^*] = \mathrm{Max}_{c(m\#)} \int U\big(Y - \pi - c^*\big(m^{\#}\big), H\big[s, m^{\#}\big]\big) f(s)\,\mathrm{d}s, \tag{11}$$

where $m^{\#}$ is defined as the solution to Equation (9). Because insurers cannot determine an individual's health state, the insurance policy cannot differentiate payments on the basis of illness severity.

An additional constraint operates on the insurance company: premiums must cover expected costs. Thus,

$$\pi = \int \big[m^{\#}(s) - c\big(m^{\#}(s)\big)\big] f(s)\,\mathrm{d}s. \tag{12}$$

The optimal insurance policy can be formally written as a problem in dynamic optimization [Blomqvist (1997)].[22] Alas, this is a complicated problem, whose algebra is

[21] Moral hazard arises, let us remember, because individuals differ in unmonitorable ways. Thus it could be on income, on health status, or on some aspect of preferences.

[22] The problem is formally analogous to the optimal tax problem in public finance when ability is unobservable [Mirrlees (1971), Diamond (1998)].

Table 5
Estimates of the optimal insurance policy

Author	Optimal policy
Feldstein and Friedman (1977)	58 percent coinsurance rate
Buchanan, Keeler et al. (1991)	\$200 deductible; 25 percent coinsurance rate
Newhouse et al. (1993)*	\$200 to \$300 deductible; 25 percent coinsurance rate; \$1,000 stop loss (assumed)
Manning and Marquis (1996)	25 percent coinsurance rate; >\$25,000 stop loss
Blomqvist (1997)**	Cost sharing declines from 27% at roughly \$1,000 of spending to 5% above roughly \$30,000

* Amounts are in 1983 dollars.
** Amounts are based on the Rand Health Insurance Experiment data.

not particularly revealing. The analytic solution balances two factors. The first is the reduced overconsumption from making people pay more out of pocket for medical care. If the coinsurance rate is increased in some range, people in that range pay more for medical care, as do people at all higher levels of spending (because their coinsurance rates have been increased). This increase boosts the efficiency of provision. Countering this, however, is a loss in risk spreading benefits. As people are made to pay more out of pocket, they are exposed to more risk, and this reduces their welfare. The optimal coinsurance rate balances these two incentives.

A small literature has simulated optimal insurance policies using this framework. Table 5 shows the results of these simulations. Table 5 reveals a wide range of disparities in optimal insurance policies. Some of the studies find that simulated insurance policies are substantially less generous than actual indemnity policies of the past 20 years [Feldstein and Friedman (1977), Blomqvist (1997)], while other studies find that they are about the same [Buchanan et al. (1991), Newhouse et al. (1993), Manning and Marquis (1996)].[23] The difference between these various estimates has not been fully reconciled, although one suspects that differing degrees of risk aversion and moral hazard are important. One suspects that real world policies will be more generous than optimal policies because of the tax distortions favoring more generous insurance: payments to insurance which are then made to providers are not taxed as income to employees, while wage and salary payments, which might be used to pay for medical care out-of-pocket, are. Indeed, other research shows that the benefits that employer health insurance policies offer are sensitive to employee tax rates [Pauly (1986)].

[23] The implication of the Blomqvist estimates for health insurance cost sharing depend on whether income losses are compensated or not.

A second important difference between real world and optimal policies is that the former almost invariably have a constant coinsurance rate, i.e., linear structures, whereas the latter do not. The optimal policy can be substantially superior. Blomqvist (1997), for example, finds that coinsurance rates should range from over 25 percent at low levels of spending to 5 percent at high levels of spending. There is likely a tradeoff between optimality and simplicity. Optimal policies can be very complicated, while real world situations are characterized by relatively simplistic structures.

If services or diseases differ in the degree of moral hazard they entail, the optimal insurance policy will differ by service or disease as well. Suppose, for example, there is a fixed number of diseases that a person can have and that moral hazard varies by disease. The insurance company can observe the disease of the person (e.g., cancer or appendicitis) but not the severity of illness within the disease. Then, the optimal insurance policy will have different cost sharing by disease [Zeckhauser (1970)]. Coinsurance formulas could just as easily depend on service (e.g., outpatient psychiatry) or locale of medical care (e.g., hospital care).[24] In practice, elasticity estimates do differ across services. The Rand Health Insurance Experience found higher demand elasticity for outpatient care than for inpatient care, and within outpatient care a greater demand elasticity for mental health care. Most health insurance policies, including Medicare, draw distinctions between services in their coinsurance schedules. Thus, Medicare has a separate hospital deductible, and private insurance plans frequently cover a fixed number of psychiatric visits.

Moral hazard is a significant concern in insurance policies but it is not one that necessarily argues for government intervention. Government insurance policies, after all, may engender just as much moral hazard as private insurance policies. There is a rationale for government to be involved in goods subject to moral hazard only if the government is better able to monitor or punish moral hazard than the private sector. This is not obviously the case in medical care.

3.2. Patients, doctors, and insurers as principals and agents

Thus far, we have implicitly assumed that patients choose the amount of medical care they want, knowing their illness, the range of possible treatments, and the prices of the treatments to them. But few patients are so well informed. In most cases of serious expenditure, it is the doctors who make the resource-spending decision, with patients and insurers bearing the costs; patients usually do not know the charge until the bill comes. Patients, physicians, and insurers are in a *principal–agent* relationship: the patient (principal) expects the doctor (agent) to act in his best interest when he is sick. Similarly, the

[24] This is analogous to the Ramsey rule of optimal taxation. The Ramsey rule states that optimal taxes on a set of commodities should be inversely related to the elasticity of demand for each commodity – in minimizing inefficiency, inelastic factors should be taxed more. The statement here is the equivalent but for subsidies instead of taxes.

insurer would like the doctor to act in its interests. Of course, patients also bear the insurance costs for seeking care, so that *ex ante* the patient's incentives and the insurer's incentives line up. But once the patient becomes sick and requires care, the parties' incentives diverge.

This three-player agency problem creates substantial problems for health insurance. To the extent that medical treatments are decided upon jointly by physicians and patients, the *supply side* of the health insurance policy (the rules about paying physicians) will matter along with the *demand-side* of the insurance policy (the rules about cost sharing for patients).

With the traditional service-benefit insurance policy, doctors and patients frequently have relatively congruent interests, which may differ from those of the insurer. Patients who face but a fraction of the costs they incur will desire excessive treatments. Service-benefit insurers usually pay more to physicians who provide more medical services. The result is that patients and physicians want essentially all care that improves health, respectively ignoring and welcoming resource expenditures. The view that physicians should do only what is best for the patient is codified in the Hippocratic Oath – providers should promote the best medical outcomes for their patients. Hippocrates said nothing about providing care the patient or society would have deemed *ex ante* to be wasteful.

Plato anticipated the application of agency theory to the health care arena by a goodly margin. He wrote that, "No physician, insofar as he is a physician, considers his own good in what he prescribes, but the good of his patient; for the true physician is also a ruler having the human body as a subject, and is not a mere moneymaker" (*The Republic*, Book 1, 342-D).[25] With the passage of 2,000+ years, fidelity to principals has slipped a bit, and new participants – insurers, government, employers, and provider organizations – have strode into the arena. But the principles are very much the same.

A more sinister view of the principal–agent problem contends that physicians manipulate patients into receiving more services than they would want, so that physicians can increase their income. This has been termed *supplier-induced demand* in the literature. An enormous amount of work in health economics has been devoted to the question of whether and to what extent suppliers induce demand. The empirical evidence on this issue is discussed by McGuire (2000). Lesson 1 notes the tradeoff between risk spreading and appropriate incentives applies on both the demand- and supply-sides of the market.

Increasingly, the arrows of responsibility among the players – who is agent, who principal – now point in all directions. For example, doctors now have responsibilities to other providers and insurers, not just to patients. Such added doctor responsibilities, primarily to hold down expenditures, ultimately enhance patient welfare, at least on an expected value basis, if not when the patient is sick. Insurers, acting for their customers as a whole, want to limit spending to only that care that is necessary; i.e., the care

[25] One might instead heed the warning of George Bernard Shaw nearly a century ago: "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" [Shaw (1911)].

patients would select given a lump-sum transfer that depends on their condition and making them pay all costs at the margin. With patients, physicians and insurers pulling in different directions, a conflict over what care will be provided frequently results.

3.3. Transactions costs

Processing claims costs money; the more claims processed the more it costs. National estimates of medical expenditure suggest that 15 percent of insurance premiums are devoted to administrative expense.[26] Someone must read the bill, approve the spending, and pay the claim. Insurance companies seek to control these costs, and policies are designed accordingly.[27]

A major part of claims processing costs – monitoring, transferring money, and the like – are invariant to the size of the claim. Size-invariant costs are a greater percentage burden for small bills than for larger bills. This suggests limiting health insurance to larger claims and having individuals pay directly smaller expenses [Arrow (1963)]. This insight gives further justification to the widespread use of deductibles and coinsurance for small bills, and for the fact that historically insurance developed first for inpatient doctor and hospital charges, where bills are the largest.

4. Relationships between insurers and providers

The medical care system is a network, with patients, monies and information flowing from one party to another. The information flow to insurers, however, is not so rich that they can guarantee that only cost-effective care will be provided. Their monitoring difficulties provide the motivation for cost-sharing in insurance policies. But cost sharing has limited value: Patients do not make the most costly decisions, the Hippocratic Oath does not extend to conserving society's resources, and risk spreading considerations severely limit what charges can be imposed.

Return now to Figure 1, the Medical Care Triad. Working solely on the left side of the triangle, the demand side, these arguments suggest that passive insurers are unlikely to be able to limit utilization appropriately. Recognizing this, insurers also work the right side of the triangle – the supply side. Increasingly, insurers attempt to provide incentives for providers to limit spending. The incentives may be imposed at arm's length, as Medicare does with its DRG system: treat a simple heart attack, and a hospital gets paid a flat

[26] This includes the expenses of paying bills as well as marketing. Divisions between these sources of administrative expense are not very precise.

[27] Of course, individuals must also bear some costs in paying bills on their own, so it is not self evident which method of payment, individually or through insurance, is cheaper. But most people implicitly assume that insurers have additional transactions costs for paying bills beyond what individuals face. Thus, there is likely to be a net transactions cost to purchasing insurance. There are also transactions costs associated with selling the policy, but they do not vary with the magnitude of claims.

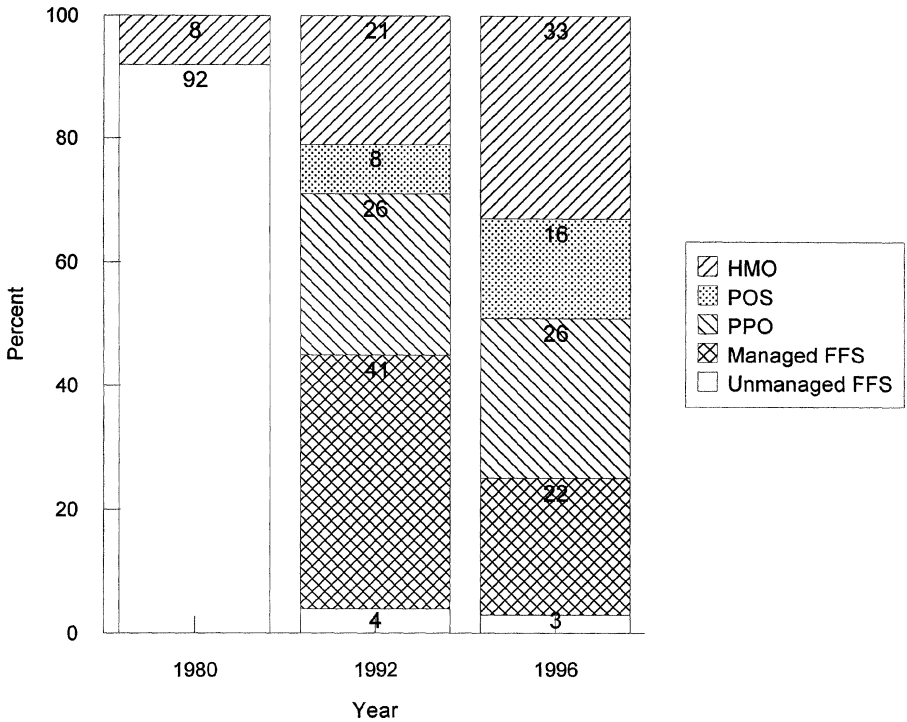


Figure 4. Changes in health plan enrollment. The sample is people with private (employer or individual) insurance. Source: Data are from Lewin-VHI.

amount, roughly $5,000. Or the insurer may form a contracting alliance with providers, as it does say with network HMOs. At the extreme, insurers and providers merge into a single entity. Uniting disparate organizations in this way enhances monitoring possibilities and better aligns incentives, but it also creates the potential for diseconomies of scope, e.g., requiring another layer of management when care is delivered.

The sweeping nature of insurer–provider interactions is indicated by Figure 4 [see also Glied (2000)]. In 1980, over 90 percent of the privately insured – i.e., employer- or self-paid – population in the United States was covered by "unmanaged" indemnity insurance. By 1996, that share had shrunk to a mere 3 percent.

Table 6 provides a taxonomy of different insurance-provider arrangements. The most limited arrangement is a "managed" indemnity insurance policy. It bundles a traditional indemnity policy with limited utilization review, for example requiring that non-emergency hospital admissions be precertified. At the most intrusive, insurers can seek to monitor care on a retail basis through tissue review committees, or on a statistical, wholesale basis by monitoring a physician or hospital's overall utilization. Such reviews can be used to refuse or reduce payment. Such intrusiveness by insurers may be unhelpful and, coming after-the-fact, may be ineffectual. It certainly is not welcome

Table 6
Key characteristics of insurance policies

Dimension	Indemnity insurance	Managed care		
		PPO	IPA/network HMO	Group/staff HMO
Qualified providers	Almost all	Almost all (network)	Network	Network
Choice of providers	Patient	Patient	Gatekeeper (in network)	Gatekeeper (in network)
Payment of providers	Fee-for-service	Discounted FFS	Capitation	Salary
Cost sharing	Moderate	Low in network; High out of network	Low in network; High out of network	Low in network; High/all out of network
Roles of insurer	Pay bills	Pay bills; Form network	Pay bills; Form network; Monitor utilization	Provide care
Limits on utilization	Demand-side	Supply-side (price)	Supply-side (price, quantity)	Supply-side (price, quantity)

to physicians. As Figure 4 shows, managed indemnity insurance, though non-existent in 1980, claimed a 41 percent share by 1992, but has fallen to 22 percent today.

Preferred Provider Organizations (PPOs), a second type of managed care, form a network of providers, including physicians, hospitals, pharmaceutical purveyors, and others, and control costs by securing discounts from them. The *quid pro quo* for the discounted fee is that insureds are steered to in-network providers. Out-of-network providers may get reduced coverage or no coverage at all. More typically, the patient's coinsurance or copayment rates are merely set lower for in-network providers. In 1991, for example, the typical PPO had an in-network coinsurance rate of 10 percent and an out-of-network coinsurance rate of 20 percent. PPOs usually impose preauthorization requirements as well, though they are rarely especially strict. As Figure 4 shows, PPO enrollment, zero in 1980, now makes up about one-quarter of the privately insured population.

Full integration creates the strongest link between insurance and provision. In the United States, these merged entities are called health maintenance organizations (HMOs). They sell their services directly to employers or individuals on an annual fee basis, and then they deliver care. There are three major types of HMOs. Within a group/staff HMO – the most common form, with Kaiser being the best known example – physicians are paid a salary and work exclusively for the HMO. The HMO may have hospitals on contract, or may run its own.

HMOs employ a range of mechanisms to limit utilization. They reflect the traditional economic instruments of regulation, incentives, and selection of types. HMOs frequently regulate physicians' practices, for example limiting the referrals they can

make or the tests they can order. But the efficiency benefits of HMOs arise much more from aligning the incentives of provider and insurer, rather than through direct regulation. Some group/staff HMO physicians are salaried; as a result, they have a weaker incentive to provide marginal care than their fee-for-service counterparts. Moreover, HMOs monitor the services that physicians provide. They may reward parsimonious resource use directly with compensation, though more likely with perks or subsequent promotion. Extravagant users are kicked out of the network. Finally, since physicians differ substantially in their treatment philosophies, HMOs can select physicians whose natural inclination is toward conservative treatment.

Given the ability of HMOs to limit utilization on the supply-side, price-related demand-side limitations can be less severe. Cost-sharing to enrollees is generally quite low – typically about $5 to $10 per provider visit, although other forms of demand-side limitation survive (for example, patients may have to get approval from their internist before seeing a specialist).

Independent Practice Associations (IPAs), or Network Model HMOs, represent a more recent innovation in managed care.[28] These plans neither employ their own physicians nor run their own hospitals. Instead, they contract with providers in the community. By limiting the size of the network, the plans secure lower costs from willing providers. In addition, these plans employ stringent review procedures. For example, patients may need approval to receive particular tests. Finally, IPAs often provide financial incentives to limit the care that they provide. For example, some plans pay physicians on a "capitated" basis. The physician receives a fixed payment per patient per year. Out of this capitated stipend, the physician must pay for all necessary medical services, possibly including hospital services and prescription drugs. The physician's incentives for cost control become even more significant when all expenditures come out of his own pocket.

In many HMOs, patients can go outside of the network and still receive some reimbursement. This is termed a Point of Service (POS) option. But reimbursement out-of-network is not as generous as reimbursement within. Use of non-network services, for example, frequently requires a deductible followed by a 10 to 40 percent coinsurance payment.

As Figure 4 shows, HMO enrollment of all forms (including POS enrollment) has increased from 8 percent of the population in 1980 (then predominantly group/staff model enrollment) to nearly half of the privately insured population today.

This vertical integration in managed care, with insurers and providers linked or united, is virtually unheard of in insurance of other types. Auto insurance, for example, is an indemnity policy. People choose what coverage they will have, what deductibles will be in force, etc. When there is a crash, the insured and the adjuster get together, perhaps at the repair shop, to negotiate the cost of the repair. The insured or the repair shop, entities having no particular relationship to the insurer, are paid that amount, less

[28] Some IPAs are older, but their form gained popularity only recently.

any deductibles, which are the responsibility of the insured. After major crashes, cost-ineffective repairs are avoided by declaring the car a total loss, giving the wreck to the insurance company and reimbursing the owner.

But such a contingent claims system could not work with health care. The claims are more frequent and uncertainties much greater, making costs much harder and more expensive to estimate. "Scrapping" a human body is rarely an inexpensive or palatable proposition. The burgeoning links between insurers and providers in health care, we believe, are a response to the *a priori* difficulty of writing contingent claims contracts in the medical sector.

Vertical integration is also important because it can elicit price discounts. Managed care partly represents a price club. In exchange for an up-front fee, the patient gets to purchase goods at a significant discount. The discounts are secured through bulk purchase bargaining, or by directly hiring the sellers. In exchange for lower prices, patients precommit to receive care from a limited set of providers, or to pay harshly for the privilege of going elsewhere.

Finally, vertical integration is important because it fundamentally transforms the principal-agent conflicts in the medical system. Physicians no longer look out for the interests of just their patients, or perhaps their patients' interests and their own. Now, physicians must watch out for the insurer as well. And patients must be more attuned to the incentives their physician is under. We note the integration of insurance and provision as the second lesson of health insurance:

> *Lesson* 2: *Integration of insurance and provision*. With medical care, unlike other insurance markets, insurers are often directly involved in the provision of the good in addition to insuring its cost. The integration of insurance and provision, intended to align incentives, has increased over time. Managed care, where the functions are united, is an extreme version. Under it, doctors have dual loyalties, to the insurer as well as to the patient.

4.1. Equilibrium treatment decisions in managed care

One can understand the impacts of managed care using a framework similar in spirit to what we described for patient cost sharing, only the physician's choices are targeted. A typical physician payment, for example, is

$$\text{Payment} = R + r \cdot \text{Cost}. \tag{13}$$

Here, R is the prospective amount and r is the retrospective amount. A fully capitated system sets $r = 0$ and $R > 0$, while a fully retrospective system sets $R = 0$ and $r \geqslant 1$. Thus, the capitated system focuses solely on incentives; the retrospective system removes all risk from the doctor.

Changing to a capitated system might affect treatment decisions in several ways. One effect is to raise the physician's "shadow price" for providing treatment – physicians might require a greater expected health benefit before providing care under managed

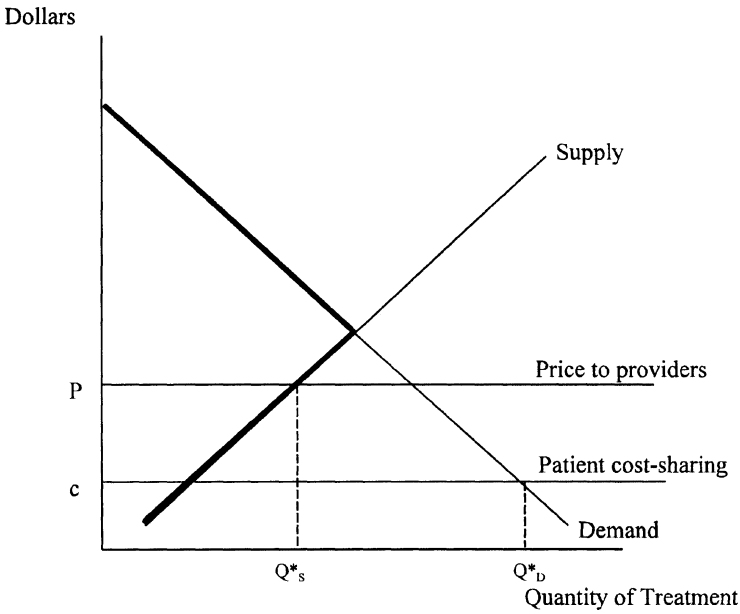


Figure 5. Conflict in quantities desired between providers and patients.

care than under traditional indemnity insurance [Frank, Glazer, and McGuire (1998), Keeler, Carter, and Newhouse (1998)]. This effect is particularly strong when the physician is capitated, and thus bears the marginal cost for providing additional care.

In addition, managed care might harmonize treatment decisions across patients. Protocols in managed care, for example, encourage or require physicians to treat patients with the same condition similarly.

In both of these circumstances, the physician's views about optimal treatment may differ from the patient's. Doctors may want to limit care while patients may want more. This divergence is particularly likely if patients pay little at the margin for medical care, as they do in many managed care plans (at least for in network services). The conflict of incentives between physicians and patients in managed care contrasts with the situation in traditional indemnity insurance, where the incentives of patients and physicians are generally aligned (although both differ from the incentives of insurers).

Figure 5 shows a potential conflict; at the prices each faces, the patient demands much more care (Q_D^*) than the physician wants to provide (Q_S^*).[29] Which level of care will ultimately be provided? Knowing how treatment decisions will be made in such an environment is difficult, as economic analysis of rationed goods in general does not

[29] We have drawn the physician's supply curve as upward sloping. This needn't be. It could be vertical or backward bending. Our point would carry through, nevertheless.

reach uniform conclusions. The situation is particularly severe in the medical care market because patients do not pay substantial amounts at the margin for medical care; thus, willingness to pay is not an accurate way of gauging individual value of services. There are several possible outcomes. One possibility is that the short-side principle, applies, which predicts that the equilibrium quantity will be the lesser of demand and supply. This is shown as the thickened line in Figure 5 and corresponds to a situation where treatment decisions in managed care are made predominantly by physicians. The short-side principle underlies much of the work on managed care [see, e.g., Baumgardner (1991), Ramsey and Pauly (1997), Pauly and Ramsey (1998)].

But the short-side principle assumes patient wishes play no role when demand exceeds supply. Treatment decisions may come out of a "bargaining" process that balances the wishes of physicians and patients [Ellis and McGuire (1986)]. One can interpret this bargaining either as an explicit process between the parties, or as the physician balancing his own self-interest (or the insurer's profits) with the best interests of the patient.

The actual level of service delivered is likely to vary with the particular medical situation. Patients with chronic conditions may know a great deal about their treatment options; the outcome may thus be close to the patient's demand. In emergency situations, the opposite may be true. The effectiveness of managed care in limiting medical spending may thus differ across settings.

4.2. Evidence on supply-side payment and medical treatment

A substantial literature examines the role of supply-side payment systems in influencing medical treatments. A change from *retrospective*, or cost-based reimbursement, to *prospective* reimbursement is typically analyzed.

Table 7 presents studies on this topic. It documents the impact of prospective payment on four aspects of hospital care: the number of admissions and transfers; average length of stay or other inputs; hospital profits; and quality of care. Prospective payment might increase or decrease hospital admissions. On the one hand, sick people might be less likely to be admitted to a hospital under prospective payment, since reimbursement for these individuals falls short of expected cost. On the other hand, hospitals might be more eager to admit healthy patients, for whom reimbursement exceeds costs. As Table 7 shows, admissions generally declined with the implementation of prospective payment.

While one might worry about whether care for the sick is excessively rationed in such a system, the literature on whether patients are being "dumped" under prospective payment (e.g., sent to public hospitals) is not particularly clear. A loose consensus is that there is some dumping of patients under PPS, but the magnitude is not particularly great [Morrisey, Sloan, and Valvona (1988), Newhouse (1989), Newhouse (1996)].

Studies examining the effect of prospective payment on average lengths of hospital stay and other inputs find nearly uniformly that average hospital stays fell with the reimbursement change. This is what theory predicts: hospitals no longer paid for each additional service will cut back on marginal care, which is expensive relative to health benefits. The effect of prospective payment on hospitals stays is uniformly strong and

Table 7
Prospective payment and medical treatments

Paper	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Frank and Lave (1986)	1981 NIMH discharge data and AHA surveys	OLS regression		LOS for psychiatric patents fell 0.3 days more in states with PPS in Medicaid		
Guterman and Dobson (1986)	HCFA in-house statistics	comparison of means and other descriptive statistics		LOS dropped 13% from 1981–84 vs. 4% in previous four years combined		
Sheingold and Buchberger (1986)	1981 PPS cost reports	simulation of provision of free care by hospitals under PPS rules				each 1% decrease in financial margin leads to 0.3–0.5% less free care provision
Carroll and Erwin (1987)	1982–85 patient records from non-random sample of 10 Georgia long-term care facilities	comparison of means				patients dying within 30 days of entering long-term care facility dropped from 14.7% to 8.3% under PPS
Feder, Hadley, and Zuckerman (1987)	1982 and 1984 AHA surveys	comparison of means			total margins for hospitals under PPS rose 2.9%, compared to no change for hospitals under TEFRA	

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Fitzgerald et al. (1987)	patients with hip fractures admitted to a municipal hospital from 1981–85	comparison of means		LOS fell from 16.6 to 10.3 days		percent in nursing home at six months after discharge rose from 13 percent to 39 percent
DesHarnais, Chesney, and Fleming (1988)	1980–85 Professional Activities Study of CPHA hospitals	comparison of means	discharges dropped 3% from 1980–85	LOS dropped 20% from 1980–85		no significant adverse effect on quality of care
Fitzgerald, Moore, and Dittus (1988)	elderly patients admitted for new hip fracture at large, mid-western community hospital,	comparison of means		LOS dropped from 21.9 to 12.6;		percent discharged to nursing home rose from 38 to 60; percent in nursing home at one year rose from 9 to 33 percent
Lave, Frank, Taube, Goldman, and Rupp (1988)	1984 Medicare PATBILL file, 1984 NIMH psychiatric discharges, HCFA, AHA, CHPS	comparison of means		LOS for psychiatric patients at PPS hospitals fell 23% under PPS; charges fell 20% under PPS		
Morrisey, Sloan and Valvona (1988)	1980, 1983–85 sample of 501 CFHA hospitals	multinomial logit and OLS for post-hospital care selection	probability of transfer increases significantly after PPS	LOS decreased in almost all major DRGs after advent of PPS		

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Newhouse and Byrne (1988)	1981, 1984–85 20% sample of Medicare claims from non-waiver states	comparison of means		LOS rose at long-term hospitals (not on PPS) relative to short-term hospitals (on PPS)		
Sloan, Morrisey, and Valvona (1988)	1980, 1983–85 sample of 501 CFHA hospitals	comparison of means		ICU/CCU days rose less in PPS states than non-PPS states from 1980–83		
Frank and Lave (1989)	1981–84 National Hospital Discharge Survey	hazard model		LOS fell 17% with per case payment for psychiatric patients		
Gaumer, Poggio, Coelen, Sennett, and Schmitz (1989)	1974–83 AHA surveys and standardized mortality rates	comparison of means				mortality rates 1% to 2% higher than predicted for urgent care patients in PPS states
Gerety et al. (1989)	Chart review of patients with hip fracture before and after prospective payment	comparison of means		LOS fell by 1.4 days		poorer discharge ambulation; no effect on nursing home residence at 1 year

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Hadley, Zuckerman, and Feder (1989)	1983–85 AHA surveys	comparison of means		LOS fell by 10.3% under PPS		
Newhouse (1989)	1983–84 5% random sample of PPS hospital bills	comparison of means; OLS regression	1/4 of cases in unprofitable DRGs move to city/ county hospitals			
Palmer et al. (1989)	patients with hip fractures admitted to a private, suburban, teaching hospital from 1981–87		LOS fell from 17.0 to 12.9 days			No effect on nursing home residence or ambulation at 6 months
Russell and Manning (1989)	annual reports of trustees of federal Hospital Insurance Trust Fund	comparison of cost projections before and after PPS			Medicare costs for 1990 reduced by $18 billion compared to projections	
Sager, Easterling, Kindig, and Anderson (1989)	1981–85 age-specific national mortality data	comparison of means				deaths in nursing homes rose by 2.6% in PPS states; no change in non-PPS states
Sheingold (1989)	1983–87 Medicare Cost Reports	comparison of means	Discharges dropped 6% in 1983 and 1984		PPS margins fell from 14.7% to 7.9% between 1983 and 1985	

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
DesHarnais, Wroblewski, and Schumacher (1990)	1980–87 Professional Activities Study of CPHA hospitals	comparison of means	admissions of psychiatric patients fell under PPS			
Folland and Kleiman (1990)	1980–85 stock market returns	seemingly unrelated regressions of excess returns			no significant excess returns to hospital management firms after PPS	
Guterman, Altman, and Young (1990)	1983–86 AHA and Healthcare Financial Management Association surveys	comparison of means			teaching hospitals have highest but fastest falling PPS margins	
Kahn et al. (1990) [series of articles using the same data in a single-volume series]	1981–82 and 1985–86 Medicare records from 297 for patients with six conditions	comparison of means	patients were 1% to 1.6% sicker at admission			no significant effects on quality: – patients admitted from home, not discharged home fell 4%; – likelihood of instability at discharge rose 22%; – patients receiving poor care fell 13%; – in-hospital mortality fell 3%; – 30-day mortality rose 1.6%

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Menke (1990)	1983–86 Medicare Parts A and B claims data	OLS regression		LOS for stroke patients fell by 2.4 days		
Ray, Griffin, and Baugh (1990)	Medicare enrollees with hip fracture in Michigan			LOS fell by 4.4 days		mortality at 1 year was unchanged
Cutler (1991)	1984, 1988 Massachusetts inpatient data	OLS regression		LOS and inpatient procedures fell under PPS		
Willke, Custer, Moser and Musacchio (1991)	1983–87 AMA Socioeconomic Monitoring System telephone surveys	comparison of means		LOS dropped by 0.6 days from 1983–87; doctors' practice hours per week rose 2.0 hours per week		
Fisher (1992)	1985–90 Medicare Cost Reports, AHA employee data	comparison of means			hospitals with Medicare profits dropped from 84.5% to 40.7% from 1985–90	
Eze and Wolfe (1993)	1982–86 Dept. of Veterans Affairs Patient Treatment Files, Medicare discharge data	ANOVA	discharges to VA hospitals rose, by 135.6% for serious cases			

continued on next page

Table 7, *continued*

Author(s)	Data	Methods	Effects of prospective payment on:			
			Admissions/ transfers	Length of stay/ other inputs	Profits	Quality of care
Hodgkin and McGuire (1994)	1983–90 ProPAC Medicare extracts	comparison of means	admissions fell 11% from 1983-90	LOS for Medicare fell and then rose from 1984–89 to levels consistent with other payers	hospital margins projected to drop from 14.5% to −10.2% under PPS	
Cutler (1995)	1981–88 New England Medicare admissions, 1981–89 Social Security death records	hazard models for readmission and mortality				compression of mortality into immediate post-admission period
Staiger and Gaumer (1995)	1984–87 25% random sample of AHA hospital file, Medicare MEDPAR discharge data	beta-logistic model of mortality				reduced payments compress mortality into period just after discharge
Ellis and McGuire (1996)	1988–92 New Hampshire Medicaid Services psychiatric discharges	simultaneous equations treatment of panel data		LOS for psychiatric patients fell 25% under PPS		
SUMMARY			admissions fall; moderate dumping from PPS to non-PPS hospitals	LOS falls significantly; other inputs fall as well	initially higher Medicare profit margins reduced over time	effects on quality ambiguous for average patient, adverse for marginal patient; lower in-hospital mortality

impressive; many studies find reductions of 20 to 25 percent over a period of 5 years or less. These studies provide among the clearest evidence that supply-side reimbursement changes do affect medical treatments.

Despite the reduction in average lengths of hospital stay, a number of studies find that profit margins fell under prospective payment. This reduction in profits came largely from a reduction in revenues. As the reduction in length of stay indicates, costs fell with the introduction of prospective payment.

In addition to examining the effect of prospective payment on quality, the literature has also examined how managed care as a whole affects medical spending. Studies of this question are summarized elsewhere, including in this Handbook [Miller and Luft (1997), Congressional Budget Office (1992), Glied (2000)]; we discuss it only cursorily here.

Virtually all studies find that managed care insurance reduces medical spending in comparison to traditional indemnity insurance. The consensus estimate would be that patients under managed care spend about 10 percent less than patients in indemnity plans, adjusted for differences in the underlying health of the two groups. The effect is somewhat greater for inpatient hospital spending, but is offset by some additional outpatient utilization in managed care insurance. Overall, therefore, incentives on the physician side clearly have an effect on overall utilization.

5. Optimal mix of demand- and supply-side controls

Given the availability of both demand- and supply-side controls, which should be employed? A first pass suggests that supply-side limitations are preferable, since providers are relatively less risk averse than are patients. In practice, however, plans with both types of limitations are sold, and indeed most plans available have a mix of demand- and supply-side cost containment features (for example, capitation with high cost sharing on out-of-network use, or indemnity insurance with utilization review).[30]

Both demand- and supply-side controls may be desirable in the presence of the other. First, patients and providers may control different features of the medical interaction. For example, the Rand Health Insurance Experiment found that patient cost sharing had a substantially greater impact on the probability that a patient uses services than on the level of services provided conditional on use [Newhouse et al. (1993)]. One can interpret this as saying that cost sharing affects insureds, but not their physicians. The evidence cited above shows that managed care can limit the level of services provided, however. An insurer or provider facing this situation might then want to combine demand- and supply-side cost sharing, the former to limit the initiation of visits and the latter to control the intensity of treatment provided within visits [Ma and McGuire (1997)].

[30] The coexisting prevalence of both types of plans may be transitional, since managed care is still relatively new. But managed care plans have increasingly been incorporating more consumer choice and cost sharing (for example, in out-of-network use). This suggests the combination is not just transitional.

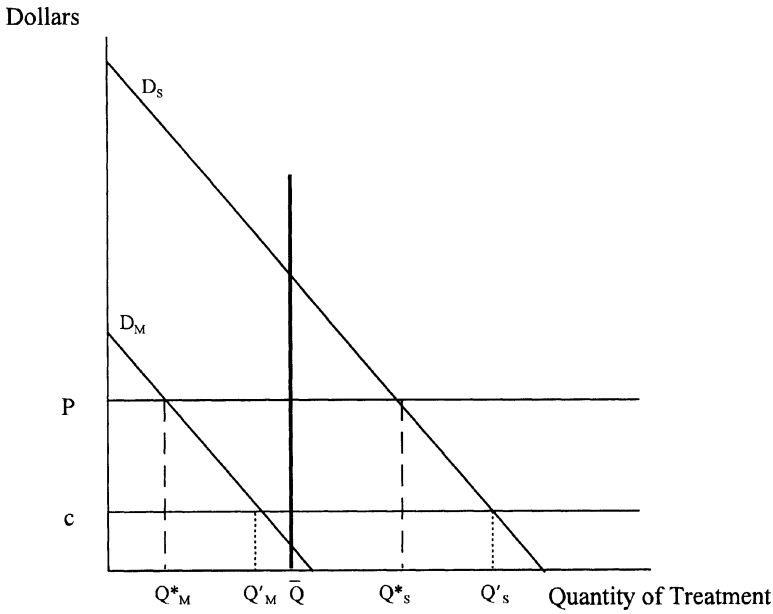


Figure 6. Demand and supply side expenditure controls.

Combining demand- and supply-side controls can also promote flexibility in types of treatment. Consider the situation in Figure 6 [Baumgardner (1991), Ramsey and Pauly (1997), Pauly and Ramsey (1998)]. There are two types of patients: those who are moderately ill (denoted M), and those who are more seriously ill (denoted S).[31] Moderately ill patients demand less medical care at any price than severely ill patients. We assume the insurer cannot distinguish the two groups, however; thus, cost sharing or quantity restrictions must apply equally to the two.

Given a price of medical care P, the optimal amounts of medical care to receive are Q^*_M and Q^*_S respectively for the moderately and severely ill. With a coinsurance policy that requires the patient to pay c for each unit of care, the equilibrium will be medical care levels of Q'_M and Q'_S. Because of moral hazard, medical care demand will be too high. Insurers might alternately adopt a fixed quantity constraint, for example $\overline{Q}$ for each patient.[32] At $\overline{Q}$, the right amount of medical care is provided in total, but not for each patient; the moderately ill patient will receive too much medical care, while the severely ill patient will not receive enough. Thus, neither demand- nor supply-side cost containment by itself yields an optimal allocation.

[31] Note that this may apply conditional on a diagnosis. For example, the conditions could be severe and moderate heart attacks.

[32] We assume that managed care features this type of restriction.

Combining demand- and supply-side cost containment can improve the situation, however. For example, starting from $\overline{Q}$, raising coinsurance will discourage utilization by the moderately ill person before the severely ill person (because the marginal value of care is much lower for the former). If the coinsurance rate necessary to deter low value utilization is small, the risk spreading loss from such coinsurance will be small, and the net welfare consequences of deterrence will be positive. The ability to limit demand by the moderately ill person, in turn, allows an increase in $\overline{Q}$, since this constraint applies only to the severely ill person. Indeed, if demand for the moderately ill person is fully constrained by the cost sharing, $\overline{Q}$ could be increased to the optimal level of care for the severely ill person. More generally, coinsurance and constraints can be combined to enable rationing in more than one dimension when there is heterogeneity of optimal treatment. A combination of the two systems may be superior to using either system alone.

A third rationale for combining demand- and supply-side controls is to limit selection behavior by providers. Providers paid on a capitated basis will have incentives to attract healthy patients and "dump" sick ones, since the provider's payment is the same with the two patients but the costs are much greater in treating the sick patient. Incorporating patient cost sharing into the insurance policy can relax the supply-side constraints and thus limit the incentives to dump patients [Ellis (1998)]. We return to this type of adverse selection in the next section.

Theoretical results to date generally suggest a combination of demand- and supply-side controls may offer significant advantages. Moreover, with so many differing incentives in the medical care system, optimal reimbursement schemes undoubtedly differ across specialties (for example, in response to moral hazard propensities) and groups of providers (for example, if the ability to bear risk differs with group size), which increases the potential for working both sides of the market. The way demand- and supply-side systems interact with each other, however, is not well understood; neither is the tradeoff between a fine-tuned system and a system that is simple and comprehensible. Real world structures suggest simplicity has its virtues. It is noteworthy, for example, that virtually all coinsurance operates at a flat rate between the deductible and any stop loss amount.[33]

6. Markets for health insurance: plan choice and adverse selection

To this point, we have talked of the design of a single health insurance plan. Most private insurance in the United States is offered on a menu basis, however, with different insureds selecting different plans. Health insurance choice is a natural way to meet differing individual preferences. Some people will prefer managed care insurance, which

[33] Simplicity and transparency may be a handicap. Conceivably insureds and doctors, not understanding what they will be respectively charged or paid for something, may behave more reasonably. For example, a low but complex coinsurance rate might be the best way to discourage utilization. It imposes less financial risk than, say, a higher flat rate, but might be just as effective in controlling use.

limits utilization but costs less, while others will opt for a more open-ended indemnity-style policy. Within indemnity insurance policies, some will be willing to bear more financial risk than others. Having these preferences reflected in market outcomes is beneficial.

In addition, health insurance choice is important to promote efficiency. Customers shopping for the lowest prices drive costs to their lowest level. Moreover, product characteristics will be shifted and new products introduced to meet consumer demands. These benefits of competition for health insurance are analogous to the benefits competition yields in other markets.

But health insurance is fundamentally different from other markets in ways that create harmful side effects from competition. The key problem is that with health insurance, unlike other services or commodities, the identity of the buyer can dramatically affect costs. Insuring a 60 year old costs 3 times as much as insuring a 30 year old, and among 30 year olds, some will have far higher costs than others. Whom one pools with in health insurance dramatically affects what one has to pay.

Generally, the sick are drawn to more generous plans than the healthy. Those expecting to use more services will, all else equal, want more generous policies than those expecting to use fewer services. If plans could charge individuals their expected cost for enrolling in each plan, the market would efficiently sort people. Such charges are generally not imposed, however, since it is widely believed that it is not fair to make people pay a lot more just because they are sick. Knowing the individual-specific prices may also not be technically feasible.

When plans can only charge average prices, generous plans will disproportionately attract sicker people, and more moderate plans will disproportionately attract healthier ones. This phenomenon is termed *adverse selection* [Akerlof (1970), Arrow (1985)]. As a result of adverse selection, generous plans will have to charge premiums above moderate plans not only because they offer more benefits but also because they attract a worse mix of enrollees. These premium differentials, if passed on to insureds, will tilt unfairly against generous plans.[34]

Adverse selection into more generous plans leads to two fundamental difficulties. First, people will choose to be in less generous plans, so that they can avoid paying for the higher costs of very sick people. Second, plans will have incentives to distort their offerings to attract the healthy and repel the sick. Since no plan would like to enroll the sickest people, *all* plans will find it profitable to distort their benefits. Indeed, even innovations that improve quality of health care may be unattractive to plans even if they come without additional cost, if they attract the wrong people. The distortion in plan provisions resulting from adverse selection is variously termed plan manipulation, skimping [Ellis (1998)], or stinting [Newhouse (1996)].

[34] This would happen, for example, if employers make a fixed dollar contribution to the premiums of each plan offered to their employees. The converse is also true; if employers heavily subsidize the difference between plan costs, employees will choose the generous plan too often.

Table 8
Benefits and costs for HIGH and LOW risk individuals

	Generous plan		Moderate plan		Basic plan	
	Benefits	Costs	Benefits	Costs	Benefits	Costs
HIGH risk	$33	$16	$20	$4	$14.00	$2.80
LOW risk	$6	$4	$5	$1	$3.50	$0.70

The consequences of these undesired side effects of competition are felt in market equilibrium. The equilibrium with adverse selection may be inefficient; it may not even exist. We express this as the third lesson of health insurance:

> *Lesson* 3: *Competition when consumer identity matters*. When consumer identity affects costs, competition is a mixed blessing. Allowing individuals to choose among competing health insurance plans can allocate people to appropriate plans and provide incentives for efficient provision. But it can also bring with it adverse selection – the tendency of the sick to differentially choose the most generous plans. Adverse selection induces people to enroll in less generous plans, so they can be in a healthier pool, and gives plans incentives to distort their offerings to be less generous with care for the sick.

Many models of adverse selection have been developed. We start with simple models and then present more advanced models.

6.1. Equilibrium with adverse selection – the basics

To understand the patterns in adverse selection, we start with the simplest possible situation [Rothschild and Stiglitz (1976), Wilson (1980)]. Assume there are two individuals, one HIGH risk and one LOW risk, and two plans, a generous plan and a moderate plan. Table 8 gives the hypothetical benefits and costs for the generous and moderate plans. We suppose that the generous and moderate plans are what HIGH and LOW respectively would design for themselves, assuming that each had to pay his own costs.[35] Note that HIGH costs more in either plan and both people use more services in the generous plan than in the moderate plan.

Equilibrium. Efficiency requires people to be in the generous plan if the additional benefits of that plan to them are greater than the additional costs they incur. In this case HIGH should be in the generous plan, and LOW should be in the moderate plan, since the additional value to HIGH of the generous plan ($13) relative to the moderate plan is greater than its additional cost ($12), while the converse is true for LOW (a benefit

[35] This assumption of respective optimality facilitates exposition, but is not required.

of $1 compared to an additional cost of $3). The efficient outcome thus separates the insureds.

Were separation to happen, the premiums would be $16 for the generous plan (the cost of HIGH) and $1 for the moderate plan (the cost of LOW). At these prices, however, HIGH would select the moderate plan; the $15 savings are greater than the $13 loss. Of course, once HIGH joins the moderate plan costs escalate, but they are still only $2.50 (the average of 4 and 1). HIGH's cost savings by enrolling in the moderate plan ($13.50) are still greater than his loss in benefits ($13). LOW will also prefer the moderate plan.

The market equilibrium will thus have both individuals in the moderate plan, a pooling equilibrium. This is not efficient, however. The reason this inefficiency arises is that individuals do not pay their own costs in each plan, but rather the average cost of the plan. Hence, HIGH mimics LOW so that he can share his costs with LOW.

There are a variety of ways to struggle back towards efficiency. Two logical candidates, assigning people to plans or charging people on the basis of expected cost, are undesirable because they respectively override free choice or sacrifice risk spreading.[36] Two additional possibilities would be to cross-subsidize the generous plan by the moderate plan [Cave (1985)], or to distort the plan offerings [Rothschild and Stiglitz (1976)].

Cross-subsidy. Suppose the moderate plan is taxed an additional $1.25 per capita, which is used to offset the premium of the generous plan. In the separating equilibrium, the premiums in the two plans will be $14.75 and $2.25, and HIGH will now prefer the generous to the moderate plan. Both insureds are better off with the subsidy than without. HIGH clearly prefers a subsidy to no subsidy. LOW also prefers the subsidy, because he pays only a $1.25 subsidy, compared to an additional $1.50 premium if he pooled with HIGH in the moderate plan.

Plan manipulation. A second mechanism to induce a separating equilibrium is to replace the moderate plan with something stingier. When faced with a stingier plan, HIGH might choose the generous plan over pooling with LOW. Making the moderate plan stingier is distasteful to LOW, but the cost to HIGH is substantially greater. This disparity in costs is what allows "hurting" the plan to produce separation.

Consider a plan called basic, also detailed in Table 8, which gives both HIGH and LOW 70 percent of the benefits and costs they would receive from the moderate plan. Thus, LOW would receive benefits of $3.50 at a cost of $0.70 were he in the basic plan and HIGH would receive benefits of $14, incurring a cost of $2.80. If basic and generous were the two plans offered, LOW would select the basic plan. If HIGH selects the basic plan as well, his premium, i.e., average cost, would be $1.75. He'd prefer the generous plan, which offers an additional $19 in benefits, but would cost only $14.25 more. LOW

[36] Partial measures are possible. For example, many employers "carve out" mental health benefits from all plans and provide those services using one insurer. Adverse selection is one rationale for this [Frank, McGuire, and Newhouse (1995)].

prefers the basic plan to pooling with HIGH in the moderate plan. Plan manipulation sacrifices efficiency, since LOW generates more net benefits in the moderate plan.

In practice, plan manipulation can take many forms. Aerobics programs, for example, will attract the vigorous healthy while spinal cord injury or high-tech cancer care facilities pull in the costly sick. There are generally more opportunities to trim a high cost-attracting service than to add aerobics equivalents.[37] Thus, we expect plans to be ungenerous with services for conditions that will predictably have high costs.

Market competition will lead to the manipulated equilibrium. Assume that the moderate and generous plans were the only offerings. All participants would pool in moderate. Introducing the basic plan would then attract LOW, HIGH would go off to generous, and the moderate plan would be abandoned.[38]

In practice, plan manipulation and cross-subsidy of premiums can be combined to promote separation. The market equilibrium will have two plans. One will be the optimal plan for HIGH, given whatever subsidy he is receiving. The other plan, which will enroll LOW, will be the plan as close as possible to moderate whose combination of subsidy and manipulation just makes HIGH prefer his optimal plan.

We show this graphically in Figure 7(a), assuming there is a continuous choice of plans.[39] We array the plans in Figure 7(a) from least to most generous – in this case variability among plans is due to differences in the percent of expenses covered. The figure shows the expected utility of LOW (the upper two lines) and HIGH (the lower two lines) at each possible level of generosity, and for both the pooling and separating equilibria. LOW does better than HIGH, since he has a lesser chance of incurring the cost of sickness. HIGH is better off pooling than separating for it allows him to shed costs; the opposite is true for LOW. For both LOW and HIGH, their optimal separating equilibrium offers less than full insurance. This might be because of moral hazard or administrative costs; without these factors each in isolation would want the most generous policy. We show HIGH as wanting full insurance in the pooled equilibrium; in our example, the benefits from the subsidy in that plan are greater than the moral hazard or administrative cost loss. In the least generous plan (no insurance), both HIGH and LOW are indifferent between pooling and separating equilibria. In the most generous plan (full insurance) the two pay the same price and get the same utility in the pooling equilibrium.

Consider the situation if HIGH and LOW are initially at A, the full insurance pooling equilibrium. An insurer that offered a plan with generosity G_1 would attract LOW,

[37] However, the Harvard University Group Health Plan – an option for Harvard faculty – offers a $50 wellness payment, which can be used say for sneakers, as an attractor.

[38] The efficiency costs of separation produced through plan manipulation may be small. That is because the moderate plan was designed for LOW. Assuming smoothness, the costs of moving away from the optimal plan are initially trivial. But the costs to HIGH, who is already far from his optimum, may be great. This disparity allows cheap distortion to produce target efficiency [Nichols and Zeckahuser (1982)].

[39] The classic diagrammatic presentation of plan manipulation (dating from Rothschild and Stiglitz) uses indifference curves. We present this in the Appendix.

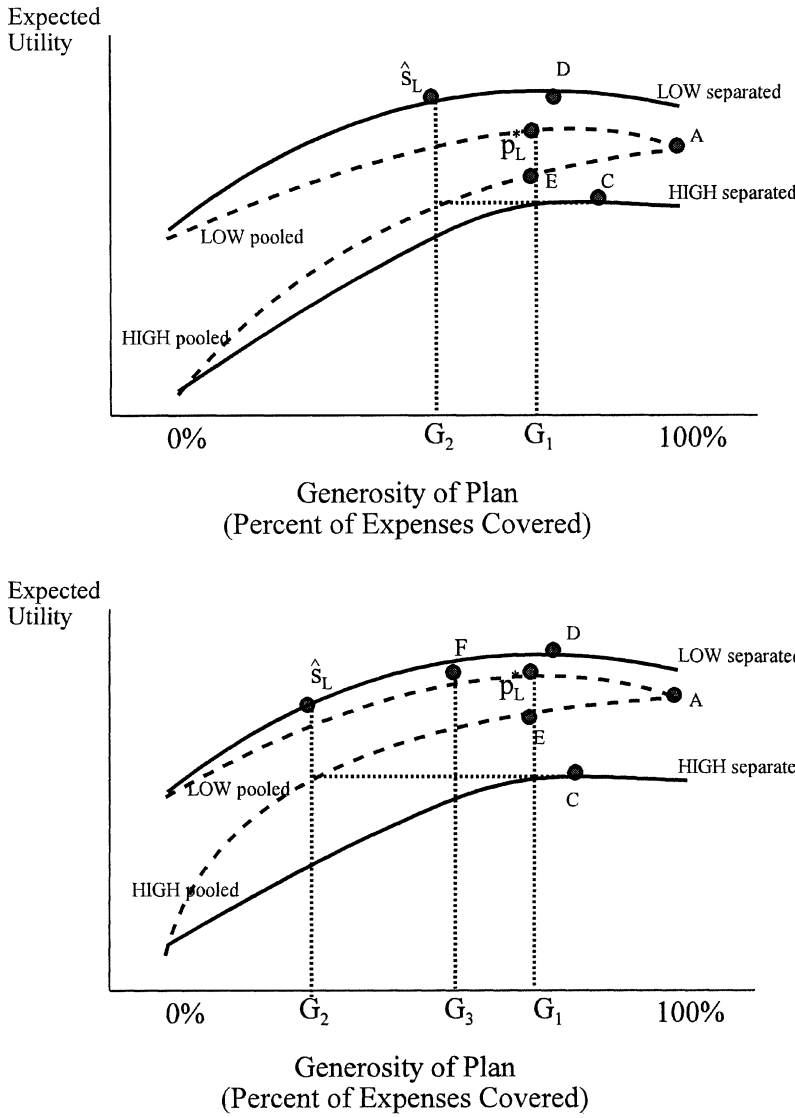


Figure 7. Reduction in insurance to separate HIGH and LOW. (a) Stable separating equilibrium. (b) Unstable separating and pooling equilibria.

since LOW prefers G_1 to A. HIGH would then move to G_1, because E is preferred to C, the separating equilibrium if only HIGH is in the generous plan. As the pooled policy becomes less generous, its attractiveness to HIGH falls. Policy G_2 makes HIGH just indifferent between pooling with LOW and the separating equilibrium at C. The

stable equilibrium will thus have two policies: LOW will be at point $\hat{s}_L$ with policy G_2 and HIGH will be at point C.

With slight changes in the curves, however, the situation at G_2 may not be stable either. Consider the situation in Figure 7(b), drawn for the case where the risk difference between HIGH and LOW is less than in Figure 7(a). Here LOW's preferred pooling equilibrium is superior to his best sustainable separating plan, $\hat{s}_L$. Thus, the separating equilibrium at G_2 will be broken by the pooling equilibrium at G_1. But the converse is also true; the pooling equilibrium at G_1 is broken by a plan say at G_3, with a price just low enough to attract LOW at F, whereas HIGH would prefer to stick with the premium and coverage at E. Once LOW went to F, however, the premium at E would have to rise, and HIGH would chase LOW to G_3. Thus, there is no stable equilibrium in Figure 7(b).

The model underlying Figure 7 assumes a frictionless world, where individuals shuttle costlessly between plans and there are no costs involved in establishing new plans. If such costs play a role, they may enable otherwise breakable equilibria to survive. For example, if establishing a plan entails high fixed costs, but individuals' transit costs remain low, p_L^* becomes stable, since breaking p_L^* with G_3 is costly but yields only temporary profits. Interestingly, greater transit costs for individuals may promote instability, since a temporary period for attracting individuals to an unstable equilibrium may last longer, and therefore be more attractive despite the fixed costs of establishing a plan. Even in this simple model, the ultimate outcome of markets with adverse selection is uncertain.

6.2. Equilibria with multiple individuals in a risk group

The simple model of adverse selection had a single HIGH and LOW risk. The lumpiness of movement implied by this specification is an important limitation of the model. With multiple individuals of a given risk type, there can also be a third class of equilibria, a "hybrid" equilibrium, to join the pooling and separating equilibria. We now show this equilibrium.

Imagine that there are now many HIGHs and LOWs, with similar tastes for insurance within each group.[40] Our example uses the parameter values from Table 8, with the \$33 benefit for HIGH under the generous plan changed to \$34. Suppose we start in the separating equilibrium, with HIGHs in the generous plan and LOWs in the basic plan. The expected utility in this equilibrium is shown by the points A and B in Figure 8. Recall that the LOWs all prefer the moderate plan to the basic plan. Imagine that they all enroll in that plan. Now suppose that instead of all the HIGHs choosing the moderate plan, only a share of them choose it. Figure 8 traces expected utility for HIGHs and

[40] A more general formulation would allow individuals within a cost class to differ on such factors as risk aversion, or in tastes for plans. Then the division of HIGHs between the moderate and generous plans would reflect the individuals' preferences.

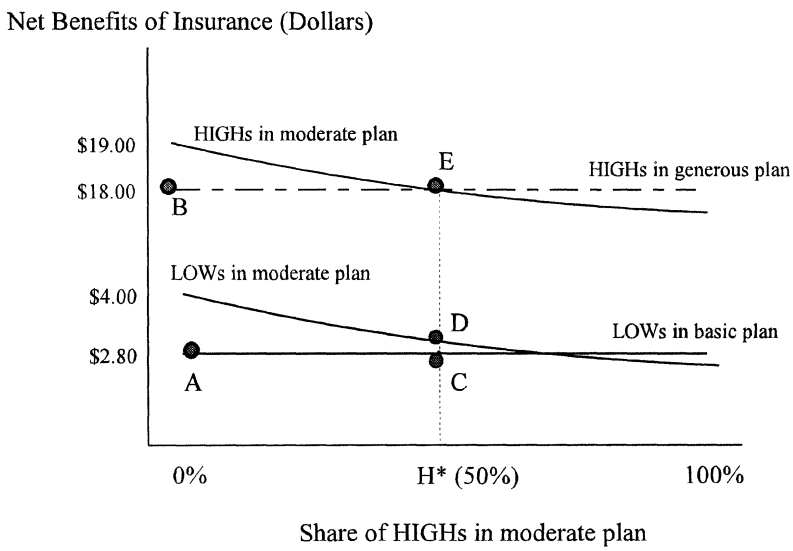


Figure 8. Hybrid equilibria with adverse selection. Note: Dashed lines assume all LOWs choose the moderate plan. The figure uses the values in Table 9, assuming the benefits to the HIGH risks in the generous plan is $34 instead of $33.

LOWs as a greater share of the HIGHs choose the moderate plan. Once H^* of the HIGHs have enrolled in the moderate plan – the number is 50 percent for our parameters – HIGHs will be indifferent between the two plans. No additional movement of HIGHs will occur.

The LOWs in the moderate plan are worse off pooling with some of the HIGHs than they would be if they had the moderate plan to themselves. But that does not indicate whether the LOWs prefer to separate themselves in basic. Indeed, in Figure 8, expected utility for the LOWs given a share H^* of HIGHs in the moderate plan (point D) is greater than expected utility in the basic plan (point C). The equilibrium with all of the LOWs[41] and a share H^* of the HIGHs in the moderate plan – what we term the "hybrid equilibrium" – is stable.

The hybrid equilibrium need not be stable, however. If the HIGHs are sufficiently costly, the LOWs will prefer the separating equilibrium to the hybrid equilibrium (point C will be above point D) and thus the two groups would separate completely.

[41] The LOWs will never end up split between the basic and moderate plan. Say the basic and moderate plans were equally attractive with a fraction of the LOWs in the moderate plan. As more LOWs moved to the moderate plan it would become more attractive. Hence, the equilibrium would tip all the LOWs into the moderate plan.

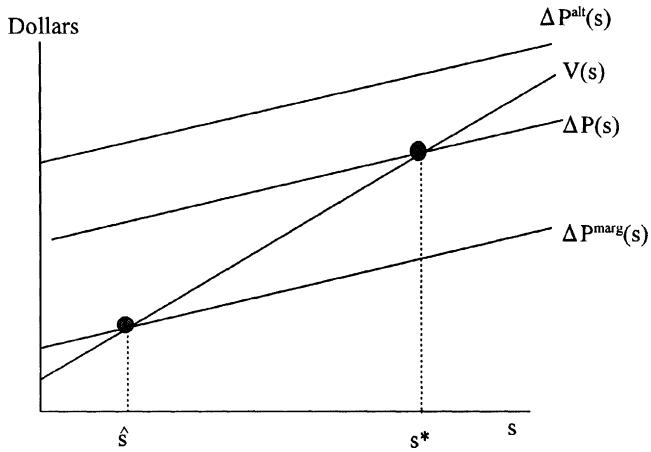


Figure 9. Enrollment consequences of adverse selection.

6.3. *Continuous risk groups*

Our two-risk-types model suggests that at least some high risks will enroll in their most preferred plan while low risks may be distorted into less generous plans. In situations with more than two risk groups, however, this situation may be reversed; the low risks may be in their preferred plans but the high risks may not. We show this using a model developed by Feldman and Dowd (1991), Cutler and Reber (1998), and Cutler and Zeckhauser (1998). The model assumes there are two pre-established plan types.

Suppose there is a continuous distribution of risks in the population, denoted by s. For simplicity, we normalize s to be the person's expected spending in the generous policy. There are two plans, one generous and one moderate. The value of more generous insurance to an individual is $V(s)$, where $V' > 0$ (the sick value generous policies more than the healthy). Figure 9 shows $V(s)$. At any additional cost for choosing the more generous policy, people will strictly divide into plans. If s^* is the sickness level of the person indifferent between the two policies, people with $s > s^*$ will choose the generous policy, whereas people with $s < s^*$ will choose the moderate policy. Average sickness in the generous policy is $s_G = E[s \mid s > s^*]$, and average sickness in the moderate policy, is $s_M = E[s \mid s < s^*]$.

Plan premiums, in turn, depend on who enrolls. We assume people in the moderate policy cost a fraction α ($\alpha < 1$) of what they would cost in the generous policy.[42] In a competitive insurance market, premiums will equal costs: $P_G = s_G$, and $P_M = \alpha s_M$. The premium difference between the two plans is therefore:

$$\Delta P(s) = P_G - P_M = (1 - \alpha)s_M + [s_G - s_M]. \tag{14}$$

[42] The literature reviewed above suggests that $\alpha \approx 0.9$ for an HMO relative to an indemnity policy.

The first term in the final expression is the cost savings the moderate plan offers to its average enrollee. The second term is the difference in the average sickness level in the two plans; it is the consequence of adverse selection.

As marginal people move from the generous to the moderate plan, the average sickness in each of the plans will rise. Depending on the distribution of s, the price difference between plans may widen or narrow. Because medical spending in practice is significantly right-skewed (Table 2), it is natural to conjecture that the premium in the generous plan will rise by more than the premium in the moderate plan. Figure 9 reflects this expectation as an upward sloping $\Delta P(s)$ curve.

The guideline for efficiency is that the price differential must be appropriate for the individual at the margin in choosing between plans. All other people would be appropriate sorted, with sicker people choosing the generous plan and healthier people choosing the moderate plan.[43] The price for the marginal individual is given by:

$$\Delta P^{\text{marg}}(\hat{s}) = (1-\alpha)\hat{s}, \tag{15}$$

where $\hat{s}$ is the person for whom Equation (15) holds. We show this schedule in Figure 9 as lying below the $\Delta P(s)$ line. $\hat{s}$ optimally delineates people in the moderate and generous plans.

Comparing Equations (14) and (15) shows that only by coincidence will the equilibrium be efficient. Suppose that the efficient allocation prevailed. From Equation (14), the price difference between the two policies will differ from this amount for two reasons. The first term in Equation (14) is generally below the efficient differential; it represents the savings from the moderate plan for the *average* person in the moderate plan, not the *marginal* person in the plan, for whom the savings would be greater. Working in the opposite direction, adverse selection (the second term in Equation (14)) will raise the premium in the generous plan relative to the premium in the moderate plan. Depending on the distribution of medical expenditures, the market differential could thus be above or below the efficient level. The right skewness of medical spending suggests that the adverse selection effect will tend to predominate. This is the situation shown in Figure 9 (by virtue of the fact that the $\Delta P(s)$ line is above the $\Delta P^{\text{marg}}(s)$ line). The premium differential for the generous plan will then be above the efficient differential, and too few people will enroll in the generous plan.

Because of adverse selection, small deviations in price can drive large differences in allocations, and indeed, the generous plan may fail to survive. Starting from $\hat{s}$, suppose the generous plan is priced too high. Marginal enrollees will depart, driving prices up still further, inducing new departures, and so on. The final equilibrium may be quite far from the efficient point. Indeed, Figure 9 also shows the possibility that the entire generous plan is depopulated. If $\Delta P^{\text{alt}}(s)$ described the cost differential, then $V(s)$ would not

[43] If preferences as well as sickness level affect the value of the generous plan, then each individual must pay his personal cost differential, $\Delta P^{i}(s) = (1-\alpha)s_i$.

intersect that line and the equilibrium would have no enrollment in the generous plan.[44] The disappearance of generous plans as a result of dynamic processes of adverse selection is termed a "death spiral". In such a situation, high risks end up in less generous plans than is optimal, while low risks get their preferred policy.

6.4. Evidence on the importance of biased enrollment

A substantial literature has examined adverse selection in insurance markets. Table 9 summarizes this literature, breaking selection into three categories: traditional insurance versus managed care; overall levels of insurance coverage; and high versus low option coverage.

Most empirical work on adverse selection involves data from employers who allow choices of different health insurance plans of varying generosity; a minority of studies look at the Medicare market, where choices are also given. Within these contexts, adverse selection can be quantified in a variety of fashions. Some authors report the difference in premiums or claims generated by adverse selection after controlling for other relevant factors [for example, Price and Mays (1985), Brown et al. (1993)]. Other papers examine the likelihood of enrollment in a generous plan conditional on expected health status [for example, Cutler and Reber (1998)]. A third group measure the predominance of known risk factors among enrollees of more generous health plans compared to those in less generous plans [for example, Ellis (1989)].

Regardless of the exact measurement strategy, however, the data nearly uniformly suggest that adverse selection is quantitatively large. Adverse selection is present in the choice between fee-for-service and managed care plans (8 out of 12 studies, with 2 findings of favorable selection and 3 studies ambiguous), in the choice between being insured and being uninsured (3 out of 4 studies, with 1 ambiguous finding), and in the choice between high-option and low-option plans within a given type (14 out of 14 studies).

Figure 10 shows a particularly salient example of adverse selection, taken from experience at Harvard University.[45] The Harvard experience is nice because adverse selection was driven by a policy change, and thus one can view the beginning of adverse

[44] Whether a death spiral actually occurs will depend on the distribution of risk levels, and the strength of the risk-preference interaction. The fatter the upper tail, the stronger the interaction, the more threatening is the possibility of a spiral. A numerical example illustrates this possibility. Suppose that the highest cost person has expected spending of $50,000 and that the average costs of the whole population in the moderate policy (with or without this person, if he comprises a small part of the total risk) is $3,000. Suppose further that the high cost person values the generous policy at $20,000 more than the moderate policy, and that he spends only $5,000 less in the moderate policy than with the generous policy (for example, a 10 percent savings if the plans are an indemnity policy and an HMO). Efficiency demands that he should be in the generous policy; the additional value of that policy ($20,000) is greater than the additional cost he imposes there ($5,000). If the high cost person were the only person in the generous policy, however, the cost of that policy would be $47,000 more than the cost of the moderate policy, which would lead him to opt for the moderate policy.

[45] See Cutler and Reber (1998) and Cutler and Zeckhauser (1998).

Table 9
Evidence on adverse selection in health insurance

Paper	Data	Empirical methods	Highlights of results	Selection
Selection between managed care and indemnity plans				
Bice (1975)	East Baltimore public housing residents (random sample)	tests of means of health status variables by Medicaid enrollment	poor health and high expected use of medical services is positively correlated with enrollment in prepaid plans; expected costs are reduced	favorable
Scitovsky, McCall and Benham (1978)	Stanford University employees' enrollment and survey data	least-squares regression of plan choice (note dependent variable is binary)	fee-for-service patients are older and more likely to be single or without children	adverse
Eggers (1980)	Group Health Cooperative (GHC) of Puget Sound's Medicare Risk Contract, 1974–76	comparison of usage statistics with control sample from Medicare 20 percent (Part A) and 5 percent (Part B) Research Discharge Files	Length of stay 25 percent higher for non-GHC patients; inpatient reimbursements per person are 2.11 times higher outside GHC	adverse
Juba, Lave, and Shaddy (1980)	Carnegie-Mellon University employees' health insurance enrollment and survey, 1976	maximum likelihood logit estimates of determinants of plan choice	lower family self-reported health status results in significantly less chance of selecting HMO enrollment	adverse
McGuire (1981)	Yale University employees' health plan enrollment statistics (random sample)	logistic regression of health plan choice given some plan is chosen	women are less likely to join the prepaid health plan than men, but no significant effect is associated with age	adverse
Jackson-Beeck and Kleinman (1983)	11 employee groups from Minneapolis-St. Paul Blue Cross and Blue Shield, 1978-81	comparison of costs and utilization for HMO enrollees and nonenrollees in period before HMO availability	HMO joiners averaged 53 percent fewer inpatient days before joining than those who chose to stay in FFS	adverse
Griffith, Baloff, and Spitznagel (1984)	physician visits in the Medical Care Group of St. Louis	nonlinear regression of frequency of visits	high usage rates at managed care plan's initiation eventually fall to lower steady-state levels	ambiguous

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Merrill, Jackson and Reuter (1985)	state employees' enrollment and utilization data from Salt Lake City and Tallahassee	tests of means in plan populations and logit regression of health plan choice	HMO joiners are younger, more often male, less likely to use psychiatric services, but have more chronic conditions in their family units	ambiguous
Langwell and Hadley (1989)	1980–81 Medicare Capitation Demonstrations	comparison of HMO enrollees and non-enrollees using two-tailed tests of means; comparison of enrollees and disenrollees using surveys	non-enrollees' reimbursements are 44 percent higher than enrollees in two years before capitation; disenrollees have worse past health	adverse
Brown et al. (1993)	Medicare spending for enrollees who stayed in traditional system versus those who moved into managed care	Comparison of spending in the two years prior to HMO enrollment	enrollees who switch to managed care had 10 percent lower spending than enrollees who stayed in traditional system.	adverse
Rodgers and Smith (1996)	summary of 1992 Mathematica Policy Research study of Medicare enrollees	measure cost differences between elderly customers covered by standard Medicare FFS and capitated HMO care	HMO patients are 5.7 percent costlier	favorable
Altman, Cutler and Zeckhauser (1998)	claims and enrollment data from the Massachusetts Group Insurance Commission (GIC)	age- and sex-adjusted analysis of costs among individuals with different plan choice histories	adverse selection accounts for approximately 2 percent of differences between indemnity and HMO plan costs	adverse
SUMMARY				adverse

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Selection of reenrollment versus disenrollment/uninsurance				
Farley and Monheit (1985)	1977 National Medical Care Expenditure Survey	OLS and 2SLS estimation of health insurance purchases	ambulatory care expenditures have an insignificant impact on health insurance purchases	ambiguous
Wrightson, Genuardi, and Stephens (1987)	disenrollees from seven plans offering different types of managed care	comparison of costs and disenrollment rates for insurees	disenrollees have lower inpatient costs and occupy less risky demographic groups than continuing enrollees	adverse
Long, Settle, and Wrightson (1988)	enrollment patterns of subscribers to three Minneapolis-St. Paul HMOs	probit estimation for chance of insuree disenrolling from each of three HMOs	likelihood of disenrollment rises significantly with increases in relative premium of own plan	adverse
Cardon and Hendel (1996)	National Medical Expenditure Survey	Tobit-style model of insurance choice	individuals who are younger, male, or in "excellent" self-reported health are significantly less likely to become insured	adverse
SUMMARY				adverse
Selection of high-option plan within type of plan				
Conrad, Grembowksi, and Milgrom (1985)	1980 random sample of claims and eligibility data for dental health insurance by Pennsylvania Blue Shield	2SLS and 3SLS estimation of demand models for premiums and total expenditures	worse self-perceived dental health corresponds to higher valuation of insurance; experience rating does not always lower premiums	adverse
Ellis (1985)	1982–83 employee health plan enrollment and expense records of a large firm	logit estimates of health plan choice	age and worse previous year's health expenses are associated with choice of more generous health coverage for the next year	adverse
Dowd and Feldman (1985)	survey data from 20 Minneapolis-St. Paul firms	tests of means of characteristics of health plan populations	fee-for-service patients are older and more likely have serious medical conditions or relatives with such conditions	adverse

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Luft, Trauner and Maerki (1985)	California state employees' enrollment and utilization data	comparisons of risk indices across plans and years	patient risk in high option indemnity and fee-for-service plans increases faster than risk in managed care	adverse
Price and Mays (1985)	Federal Employees Health Benefits Program proprietary data	comparison of costs and premiums across plan choices	high option Blue Cross plan undergoes a premium spiral with enrollment cut in half over only three years	adverse
Marquis and Phelps (1987)	Rand Health Insurance Experiment	probit estimation for hypothetical take-up of supplementary insurance	families in highest expenditure quartile were 42 percent more likely to obtain supplementary insurance than those in lowest quartile	adverse
Ellis (1989)	claims and enrollment data from a large financial services firm	analysis of different plans' member characteristics and expenses	employees in high option plan are 1.8 years older, 20.1 percent more likely to be female, and have 8.6 times the costs of the default plan.	adverse
Feldman, Finch, Dowd and Cassou (1989)	survey of employee health insurance programs at 7 Minneapolis firms	nested logit for plan selection	age varies positively with selection of a (relatively generous) IPA or FFS single-coverage health plan	adverse
Welch (1989)	Towers, Perrin, Forster, and Crosby Inc. study of Federal Employees Health Benefits program	comparison of premiums between high and low option Blue Cross plans for government workers	high-option premium is 79 percent higher than low option	adverse

continued on next page

Table 9, *continued*

Paper	Data	Empirical methods	Highlights of results	Selection
Marquis (1992)	plan selection of families in Rand Health Insurance Experiment	comparison of plan choices with age/sex adjustments under various group-rating regimes	73 percent more individuals in high risk quartile choose most generous plan than those in low risk quartile, even with age/sex/experience rating	adverse
Van de Ven and Van Vliet (1995)	survey and claims data from 20,000 families insured by largest Dutch insurer, Zilveren Kreis	regression of risk factors on prediction error of difference in costs between members of high- and low-cost plans.	age-and sex-composition of plans explain 40 percent of error in predicted cost differential between plans	adverse
Buchmueller and Feldstein (1997)	University of California Health Benefits Program enrollment figures	historical analysis of enrollment changes and premium increases	two high-option plans suffered fatal premium spirals in a six-year period; a third was transformed from FFS into POS to prevent a spiral	adverse
Cutler and Reber (1998)	claims and enrollment data from Harvard University	calculation of welfare loss and simulation of long-run effects of changes in health plan prices	adverse selection creates a welfare loss equal to 2 percent of baseline health spending; price responses in long run are triple those in short-run	adverse
Cutler and Zeckhauser (1998)	claims and enrollment data from Harvard University and the Massachusetts Group Insurance Commission (GIC)	analysis of different plans' member characteristics and expenses	employees in GIC's FFS plan spend 28 percent more, are older, and have significantly more births and heart attacks than HMO members	adverse
SUMMARY				adverse

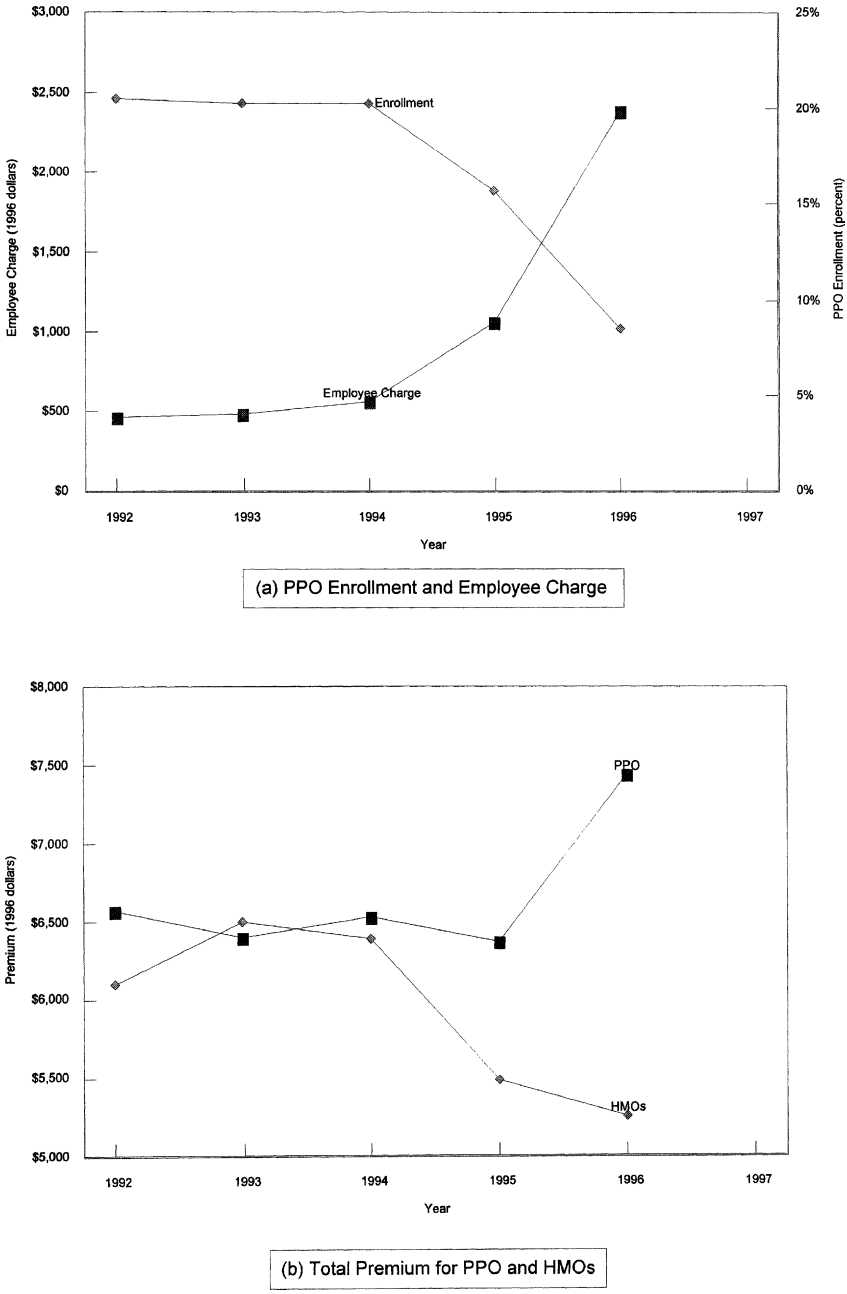


Figure 10. Adverse selection at Harvard University. Note: Dollar figures are for a family policy. Source: Cutter and Reber (1998).

selection and its subsequent effects. In the early 1990s, Harvard University offered its employees two types of health insurance plans: a generous PPO and a number of HMOs. The University paid about 90 percent of each plan's premium; thus, the employee cost of the PPO, shown in Figure 10(a), was a relatively modest $500 per year. To trim costs, Harvard in 1995 moved to a more competitive health insurance system. Under the new system, the University pegged its contribution at a fixed percentage of the lowest cost plan. Employees paid the entire amount above this for the plan of their choice. The hope was that competition among plans would drive down premiums and thus save the University money.

When the new system was introduced, the cost of the PPO rose, and PPO enrollment fell. As Figure 10(a) shows, about one-quarter of PPO enrollees left the plan between 1994 and 1995. These enrollees were disproportionately the younger and healthier employees in the PPO, however. As a result of the biased disenrollment, the PPO lost money in 1995; in 1996, it had to raise its premium by nearly $1,000. This led to a further decline in PPO enrollment; over half the remaining PPO enrollees left the plan after 1996. Again, these employees were disproportionately younger and healthier than those that remained in the PPO. Thus, the PPO premium lost money again in 1997 and would have had to increase premiums substantially in 1998, just to prevent losses. In fact, the required premium increase would have been too large for the insurer and Harvard to bear. The PPO was disbanded before that year. Adverse selection thus produced a death spiral, and did so very quickly. The disappearance of the PPO is a welfare loss to employees who would have chosen it at their individual-specific cost. Cutler and Reber estimate the size of the welfare loss at 2 to 4 percent of baseline premiums.

The importance of adverse selection has had direct impacts on policy. For example, Brown et al. (1993) found that Medicare enrollees who enroll in a managed care plan would have spent 10 percent below average if they had been in the traditional system. Since Medicare paid only 5 percent less to managed care companies for enrolling these people, Medicare lost money as HMO enrollment increased. In 1997, Federal legislation reduced payments to HMOs by an additional 5 percent, to avoid these continuing losses.

6.5. Evidence on the importance of plan manipulation

There are substantially fewer empirical studies on plan manipulation than on adverse selection. Plans, of course, differ greatly in their generosity. But it is difficult to know, and plans do not want to reveal, the extent to which the observed variation in plan benefits reflects manipulation by the plans to attract healthy risks as opposed to the self-interested choice of insurance arrangements among people already enrolled in the plans. Adverse selection aside, plans with sicker enrollees probably should be more generous.

Though evidence on plan structures is ambiguous, the marketing of managed care plans shows clear efforts to promote favorable selection. Maibach et al. (1998) document the marketing practices managed care plans use to attract healthy Medicare enrollees, including television ads that show seniors engaged in physical and social ac-

tivities and marketing seminars held in buildings that were not wheelchair accessible. Whether such practices extend to the types of benefits these plans offer is unknown.

6.6. The tradeoff between competition and selection

In weighting the consequences of competition, losses from adverse selection must be balanced against the gains, if any, from lower premiums that competition induces. The Harvard University study discussed above [Cutler and Reber (1998)] shows such a tradeoff. As Figure 10(b) demonstrates, premiums for the HMOs fell by over $1,000 when the University moved to flat-rate pricing. The savings to Harvard from these lower premiums was estimated at 5 to 8 percent of baseline health spending. This cost savings is greater than the 2 to 4 percent loss from adverse selection noted above. Thus, the net effect of competition in the Harvard circumstance appears to be beneficial, although the adverse selection losses were quite large.

With few exceptions [Wholey et al. (1995), Feldman and Dowd (1993), Baker and Corts (1995)], few studies have examined how competition affects health insurance premiums. It is often difficult to gather data on premiums, since most insurers charge different groups different amounts. In addition, premiums need to be adjusted for differences in the quality of benefits, but the many dimensions of quality are very difficult to control for. Thus, the tradeoff between cost savings and adverse selection in other situations is generally unknown.

6.7. Risk adjustment

The fundamental question about health insurance design is how to achieve the benefits of competition while containing the costs of adverse selection. A natural solution is suggested by the model above. Suppose that individuals were not charged the full difference in premiums between plans, but that instead the employer or government entirely running the insurance system offset some of the difference. For example, if the generous plan has above average risks in the amount $E[s|s > s^*] - E[s]$, the government would give the plan a per capita subsidy equal to this amount. The subsidy would be financed by a tax on the moderate plan, which has below average risks, by the amount $E[s|s < s^*] - E[s]$. The contribution from the plans would just match,[46] so there would be no net cost to the government.

In a competitive market, plans that receive subsidies (or are forced to pay taxes) would pass these subsidies on to consumers. Therefore, the premium for the generous plan would fall to $P_G = s_G - \text{subsidy}_G = E[s]$, and the premium for the moderate plan would rise to $P_M = \alpha s_M + \text{tax}_M = E[s] - (1-\alpha)s_M$. The adjusted premium difference between the plans, which individuals would face, would thus be

$$\Delta P^{\text{adj}} = P_G - P_M = (1-\alpha)s_M. \tag{16}$$

[46] This is because, taking expectations, $(E[s|s > s^*] + E[s|s < s^*])/2 = E[s]$.

This quantity is the savings for the *average* person in the moderate plan. It is closely related to the optimal price difference in Equation (15), which is the savings for the *marginal* person in the moderate plan. Plan choices made on the basis of the price difference in Equation (16), though not optimal, are likely to be more efficient than plan choices made on the basis of unadjusted price differences.

This form of differential payment is termed "risk adjustment" [Van de Ven and Ellis (2000)]. Risk adjustment must be carried out by some entity that can require individuals to insure or convince them to do so through subsidies. Otherwise, low cost individuals would choose not to participate. One possibility would be for the government to impose risk adjustment, whoever is the payer. But employers providing subsidized health insurance can do the job just as well. Employers have an incentive to risk adjust since it promotes efficiency and thus lowers the overall cost of providing health coverage.[47]

Empirically, risk adjustment can be carried out in four ways. Plans can pay or receive payments based on: (1) demographic variables (for example, more for taking on older people); (2) medical conditions (for example, more for people with diabetes); (3) past medical expenditures, which help predict future expenditures; or (4) actual experience in a year (for example, $50,000 extra for each organ transplant patient). The first three approaches attempt to predict experience; the last is after-the-fact reinsurance.

The tradeoffs between these different forms of risk adjustment are related to the ability of health plans to manipulate the risk adjustment system. Information about diagnosis, past claims, and actual use increase the ability to measure differential enrollment, but are susceptible to distortion by the plans. For example, plans may code borderline people as having diabetes if risk adjustment is done on the basis of the number of diabetics. Plans might creatively assign costs to high cost cases, when such cases are largely reimbursed. Even if risk adjustment is done on a prospective basis, plans have an incentive to exaggerate current sickness and expenditure levels, since the vast majority of insureds stick with their plans from year to year. A final, at least theoretical, concern about risk adjustment is that it may diminish plans' incentives to maintain their enrollees' health. Keeping people healthy disqualifies the plan from receiving additional risk adjustment payments, thus reducing the value of the health investment.

Because so few employers or governments have used formal risk adjustment systems, the relative advantages and drawbacks of different risk adjustment methodologies are unknown. New efforts may provide some of this information, however. In January 1999, in a major initiative, the federal government announced its intention to employ risk adjustment on the basis of past diagnoses to pay HMOs that enroll individuals in Medicare. Evaluating the impact of this system is a major research priority.

[47] Some employers have made second-best efforts to implement risk adjustment, at times inadvertently. The heavy subsidy of premiums – many employers pay 85 percent or more – in effect covers 85 percent of cost differentials due to varying mixes of insureds. Alas, heavy subsidies also significantly diminish the incentives of insureds to shop around, hence of health plans to hold down their costs.

7. Person-specific pricing, contract length, and premium uncertainty

Adverse selection is a problem of asymmetric information – individuals know their likely medical care utilization but insurers either do not, or are not allowed to use this information. Increasingly, however, information is becoming equalized. Insurers question individuals or monitor their past utilization to forecast their future costs. Equipped with such knowledge, insurers may know more about expected costs for the groups they are insuring than the members of the groups do themselves.

Insurers can use this information to set premiums. While such "experience rating" is rare at the individual level, it is common at the group level. Most private health insurance in the United States is at least partly experience rated. The bigger the group purchasing insurance, the more likely is experience rating. Hence, older and sicker groups are charged more per capita for the same coverage.

But experience rating creates its own set of problems, particularly when carried out at the individual level. When people face premiums that depend on their sickness, they are denied a form of insurance – the ability to obtain the same insurance premiums as their peers at the same price. The welfare loss can be significant.

Consider, for example, a situation where individuals are insuring themselves, diabetes is the only disease, and both people and plans know who is diabetic. Plans would offer full insurance to everyone but would charge diabetics more than non-diabetics; after all, no one who is not diabetic would be willing to pay extra to insure the diabetics. Given the distribution of diabetics and non-diabetics, the higher premiums charged to diabetics create a distributional issue. Diabetics pay more, and non-diabetics pay less relative to level premiums.

But from an *ex ante* perspective, before anyone knows who will contract diabetes, the distributional issue represents an efficiency loss. Suppose that before an individual knew if she would be diabetic or not – potentially before birth – she was offered insurance against the risk that she would become diabetic and thus face higher insurance premiums in the future. Full insurance would guarantee that if she developed diabetes, the policy would give her sufficient income each year to cover the higher diabetes premium she would then face. The benefits would be financed by payments from non-diabetics. Individuals would be willing to purchase this insurance were it sold at fair odds; they get a reduction in financial risk at no expected cost.

In real-world markets, however, such insurance against falling into a worse risk class is not offered. Some of the insurance would have to be purchased before birth. People obviously cannot do this, and even their parents might be unable to buy it for them, if there is a genetic predisposition towards disease. Other insurance could wait until mid-life for the unpredictable infirmities of old age. The key is to contract for insurance before the risk is resolved. While long-term anticipatory insurance is possible, health insurance in actual markets is rarely sold for over one year. People consequently lose welfare *ex ante*; there is an insurance policy they want but cannot obtain.

This loss at first may seem counterintuitive: everyone has full information and everyone gets full insurance every year. Where is the source of the loss? The welfare

loss derives from a missing market for insurance against one's risk type. Risk-averse individuals would like to insure against the possibility of being discovered to be high risk. There is no market where they can do so, however. Given that a market is missing, there is no guarantee that efficient pricing on the basis of known information as opposed to level pricing (as if ignorant) will enhance welfare. This illustrates the theory of the second-best. The market failure might also be thought of as a recontracting failure. We recontract for health insurance annually despite the fact that we learn about expected future health costs during the year. Such periodic recontracting breaks the contractual arrangements that would characterize optimal insurance.

This problem has variously been termed the problem of renewable insurance or the problem of intertemporal insurance [Pauly, Kunreuther, and Hirth (1995), Cochrane (1995), Cutler (1996) and Zeckhauser (1974)]. It is likely to grow in importance in health insurance markets as our ability to predict medical spending rises, as it will, for example, through advances in genetic screening. We note this as the fourth lesson of health insurance:

> *Lesson* 4: *Information and long-term insurance*. More information about individual risk levels allows for more efficient pricing of risk, but portends a welfare loss from incomplete insurance contracts.

Might markets develop to deal with this problem? Some possibilities suggest themselves. People might purchase insurance for their lifetime rather than annually. If insurance choices were made early enough (or high-cost people were compensated when insurance choices were made), people would not suffer from knowledge gained over time. Long-term purchases, such as those associated with whole life insurance, are made in this fashion. Individuals buy level premium life insurance when they are young and healthy; they will wish to retain it, even if relatively healthy, when they grow old and annual risks escalate.

In theory, health insurance could be sold for the long term on a level premium basis. In practice, matters will be more complex. Much health insurance is now bundled with the provision of care. If an individual left a geographic region, he might have to change provider, and no new provider/insurer would want to take him own at his old level rate. Portability is but one problem. Once individuals purchase lifetime medical insurance, why should an insurer strive for efficiency when people are stuck in his plan? This problem is exacerbated since the insurer must agree to pay for or provide a changing level of services. Health insurance policies optimally change from year to year, as medical technology improves and knowledge about optimal treatments expands. Finally, with future medical costs so unpredictable, insurers cannot take on the risk, which would apply to all policies, that costs will escalate beyond expectation. With life insurance, by contrast, portability, changing service mix, and varying costs are not problems.

A second approach to long-term health insurance would be to develop policies offering insurance against learning one is high cost [Cochrane (1995)]. Imagine that people purchase two insurance policies in a year; one to cover their medical costs that year, and a second to cover any increase in premiums they may face in the future. The second

policy – the "premium insurance" policy – might look like a standard health insurance policy: people pay in money and if they learn they are likely to have high costs in the future they receive money back. Full premium insurance would give people an amount of money equivalent to the discounted expected increase in their future medical spending they learn about during the year.[48] Why don't we observe premium insurance? Several factors have been identified. Moral hazard (people with premium insurance would take insufficient care of their health) and adverse selection (people expecting declines in health would more likely take up the insurance) are possibilities.

The aggregate risk phenomenon provides still a third explanation [Cutler (1996)]. Full premium insurance would have to insure a person against the risk that the medical policy that a representative individual will need in the future will cost more then than it is forecast to cost today. But future medical costs are not known. For example, a half century ago, the cost of treating cardiovascular disease patients was minimal with little prospect for rapid increase. Bypass surgery, angioplasty, and the like unexpectedly increased the cost of treating cardiovascular disease. Diversifying such a risk of significant cost increases for a common ailment is not possible. It is what is termed an aggregate as opposed to an idiosyncratic risk, where the latter apply to individuals one at a time. Insurers generally eschew aggregate risks. By contrast, insurers accept risks readily when they can lean comfortably on the Law of Large Numbers to spread them, as they can with idiosyncratic risks. They generally refuse to write insurance for risks that are unpredictable or nondiversifiable since they could bankrupt the company. Cost increases associated with future medical care suffer both disqualifications.

The result is that even though improved insurer information may reduce adverse selection over time, problems in insurance markets may grow. If people are increasingly charged on the basis of their individual risk characteristics, the efficiency losses could be severe.

Does employer-based insurance, where individuals choose from a menu of options, help? Under such plans, there is a range of potential costs individuals can face for choosing more generous insurance. At one extreme such plans are fully subsidized; people pay the same amount for each plan. At the other extreme there is no subsidy; people pay the expected cost in each plan on a group or individual basis. A system of risk adjustment lies in between; people pay the average cost of more generous plans assuming the mix of insureds is constant across plans.

We have stressed the efficiency aspects of risk-adjusted premiums, but such a system may not spread risks to a sufficient extent. Even in the perfect risk-adjusted equilibrium, the sick will pay more than the healthy, since they will be more attracted to the generous plan. People would presumably like to insure some of even this efficient price difference. There is, in terms of our earlier discussion, a tradeoff between moral hazard and risk sharing. Risk spreading considerations suggest that people should pay nothing

[48] This is related to the solution in Pauly, Kunreuther, and Hirth (1995). They propose paying a large premium in the first year, which is used to finance additional care for those who become sick in later years.

additional for selecting more generous plans, assuming risk level was the driving factor in their choice. Efficiency dictates that they should pay the expected additional cost they incur by choosing more generous care. The optimal differential lies between the two extremes, at the point where the marginal costs in terms of misallocation of people across plans exactly offsets the marginal benefits of increased risk sharing. Of course, price setting to this level of refinement may not be possible.

8. Insurance and health outcomes

Our empirical analysis to this point has focused on the impact of health insurance on medical spending. Ultimately, people care about health insurance because they are concerned about their health. A central research issue is therefore how alternate insurance arrangements affect health.

Much policy rhetoric expounds on the effects of not having insurance on health. Evidence on this issue shows that the effect of being without insurance can be large. See Weissman and Epstein (1994) for a review. For our purposes, however, we are interested in how variations among the set of insurance plans affect health. One might expect an attenuated version of the same finding – that people carrying less generous insurance, either indemnity insurance with high cost sharing or managed care insurance, would suffer worse health outcomes than people with more generous insurance. This might be particularly expected since medical treatment differs across insurance categories.

But several factors work in the other direction. Some of the additional care provided under more generous insurance may be iatrogenic (harmful to the patient), conceivably provided by physicians to increase their income. Perhaps more important, managed care policies may improve outcomes. One feature of managed care is that it standardizes the care that is received by classes of patients. These standards, if based on sound science and carefully crafted to patient characteristics, may be superior to what physicians conclude on their own. In addition, managed care usually involves less cost sharing for primary care, preventive services, and prescription drugs than does indemnity insurance. Greater use of these services may improve health outcomes.

Evidence on the effect of different insurance arrangements on health outcomes generally suggests very little difference in health produced across plans. The clearest findings on the impact of differing levels of demand-side cost sharing emerge from the Rand Health Insurance Experiment [Newhouse et al. (1993)]. The Rand study measured a broad array of health indicators. For most people, outcomes did not differ across plans. This is true even though spending differed across plans by up to one-third. Insurance did have a small effect on the health of the sick poor: poor people achieved better outcomes in more generous plans with blood pressure control, vision correction, and filling decayed teeth. Of course, the Health Insurance Experiment lasted for only a few years, which may have tilted the test against more generous plans. Increased primary and preventive care, even if strongly beneficial, may not be so important in such a short period of time.

Many studies have examined the impact of supply-side cost sharing on medical outcomes. Such studies must adjust for differing population mixes across plans, which is a difficult challenge. Important evidence comes from the implementation of prospective payment for hospital admissions covered by Medicare. At the time of the change, the critics of the new prospective payment warned that patients would be discharged from hospitals "quicker and sicker." Several papers examined this question, as shown in Table 7. The most detailed studies are the papers grouped under Kahn et al. (1990), which examined patient medical reviews before and after prospective payment was implemented to measure changes in health. That research found no increase in adverse outcomes for the average patient after prospective reimbursement, although it did find that with prospective payment more patients were discharged from the hospital in an unstable condition. The lack of significant adverse effect on quality of care was also found by Desharnais, Chesney, and Fleming (1988).

Some papers have found evidence of adverse outcomes resulting from prospective payment. Fitzgerald et al. (1987, 1988) found that patients admitted to a hospital in the midwest with a hip fracture were discharged sooner after prospective payment but were more likely to be in a nursing home 6 months and 1 year after the hip fracture. In response, many other researchers have examined this question, finding that length of stay for hip fracture patients fell but there was no effect on nursing home residence, functional status, or mortality after 1 year [Gerety et al. (1989), Palmer et al. (1989), Ray, Griffin, and Baugh (1990)].

Two studies have looked at the impact not of the prospective payment system, but of the revenue changes stemming from prospective payment [Cutler (1995) and Staiger and Gaumer (1995)]. These studies compared patients admitted to hospitals that lost revenue with patients admitted to hospitals that gained revenue. The former patients experience a compression of mortality into the period just after the hospital admission in comparison to the latter; some classes of patients that formerly survived several months after being hospitalized did not live as long after revenues fell. The effect diminished over the succeeding year, however. For patients who survived a year or longer, there was no increase in mortality.[49] The authors conclude that price changes have a small adverse effect on the very sick, but little effect on others.

A second set of evidence examines the effect of managed care on health. Miller and Luft (1997) summarize 35 studies comparing medical outcomes in managed care and indemnity insurance. They find no clear difference; some studies find that managed care does worse, while an equally large number find it does better. Many find no difference in outcomes.[50]

One is tempted to conclude from these findings that managed care is superior to traditional insurance – it saves money without substantial adverse effects. Such a conclusion

[49] After a phase-in period, hospital payments in total were not substantially affected by prospective payment, so these results are consistent with the Kahn et al. (1990) finding of no change in health for the average patient.
[50] See also Cutler, McClellan, and Newhouse (1998).

is premature, however, until long-term evidence on the effect of managed care has been obtained. We note the focus on health and lack of conclusive results as the fifth lesson of health insurance.

> *Lesson* 5: *Health insurance and health*. The primary purpose of health insurance and delivery is to improve health. Unfortunately, conclusive results are not in on which insurance and provision arrangements do this most effectively.

9. Conclusions and implications

Health insurance has a complex anatomy. The lens of economics brings many of its critical features – incentives, risk spreading and asymmetric information – into sharp focus. The understanding thus gained, however helpful, does not solve all of the problems. Indeed, the primary message of this chapter is that health insurance design is a challenging exercise in the second-best. On each of a variety of dimensions, goals must be traded off against each another, since first principles are in conflict.

Our lessons about health insurance, highlighted in Table 10, are instructive in this respect. We start with a single insurer. Lesson 1 stresses the tradeoff between efficient risk spreading and excessive utilization. Optimal risk sharing puts all the burden on the risk-neutral insurer, but this induces moral hazard (excess consumption of services) and possibly supplier-induced demand (excessive provision). Lesson 2 finds that integration of insurance and provision of services, which is absent in other insurance contexts, may be desirable to align producer and insurer incentives in the delivery of medical care.

Lessons 3 and 4 highlight second-best problems in the health insurance marketplace. Lesson 3 shows that competitive markets, the traditional lodestar of economics, may have undesirable side effects in health insurance. Most important, competition induces adverse selection, hence the misallocation of people to plans and the incentive for insurers to trim their offerings to deter the sick. In theory at least, risk-adjustment methods, which are just now being tried in practice, can counter these phenomena. Lesson 4 alerts us, however, that even if we slay the dragons of adverse selection and plan manipulation, a fierce risk remains. Since insurance is written on an annual basis, individuals are denied crucial protection against becoming sick and having their premiums escalate substantially in the future.

Lesson 5 reminds us that the ultimate goal of health insurance does not involve the usual economic concepts of prices, incentives and costs. Rather, the central objective of health insurance is to maintain and enhance our health. The payoff question, therefore, is what can we get for alternative levels of expenditure? The contribution of economics is to enable us to sketch the production function.

Health insurance is a service in society, like a haircut or tennis lesson. Why then does health insurance cause so many more problems than the other two? Both the insurance aspect, and its area of application, health, produce problems. In any insurance situation, moral hazard and adverse selection plague outcomes. In the case of health insurance, the problems are magnified, since health-promoting and care-seeking actions are difficult

Table 10
Five central lessons about health insurance

Lesson 1: *Risk spreading versus incentives*	Health insurance involves a fundamental tradeoff between risk spreading and appropriate incentives. Increasing the generosity of insurance spreads risk more broadly but also leads to increased losses because individuals choose more care (moral hazard) and providers supply more care (principal-agent problems).
Lesson 2: *Integration of insurance and provision*	Medical care is unlike other insurance markets in that insurers are often involved in the provision of the good in addition to insuring its cost. The integration of insurance and provision, intended to align incentives, has increased over time. Managed care, where the functions are united, is an extreme version. Under it, doctors have dual loyalties, to the insurer as well as the patient.
Lesson 3: *Competition and consumer identity*	When consumer identity affects costs, competition is a mixed blessing. Allowing individuals to choose among competing health insurance plans can allocate people to appropriate plans and provide incentives for efficient provision. But it can also bring with it adverse selection – the tendency of the sick to prefer the most generous plans. Adverse selection induces people to enroll in less generous plans, so they can be in a healthier pool, and gives plans incentives to distort their offerings to be less generous with care for the sick.
Lesson 4: *Information and long-term insurance*	More information about individual risk levels allows for more efficient pricing of risk, but portends a welfare loss from incomplete insurance contracts.
Lesson 5: *Health insurance and health*	The primary purpose of health insurance and delivery is to improve health. Unfortunately, conclusive results are not in on which insurance and provision arrangements do this most effectively.

to monitor, and it is widely believed to be unfair to charge people more if they contract diseases that are not their fault. Moreover, the payoff from health insurance, unlike say life insurance, is quite variable, and subject to human choice made after the contract is written. In addition, for justifiable reasons, health care is written on an annual basis, though today's chance outcomes often have cost implications that stretch for decades. Finally, health has a privileged position above other goods and services. For a range of philosophical and moral reasons, societies care deeply that their citizens receive health care, even if that is not what they would buy were they given the money.

These fundamental issues surrounding the equitable and efficient provision of health insurance make government involvement inevitable, and in many contexts desirable. The range of government involvement in health care and health insurance is enormous. At one end, many governments provide medical care directly; they raise money through

taxes, hire doctors and run public hospitals. Less extreme are countries where the government is the sole insurer, but provision of services remains private. More market-oriented systems such as the United States have most of the population in private insurance and most of the provision of medical care done by private providers. Even there, though, government plays a sizeable role, refereeing the playing field and insuring those who the market would leave behind. Thus, the federal government insures people through Medicare and Medicaid, provides tax subsidies to private insurance, defines permissible structures for supplementary Medicare insurance, and requires insurers to cover people who recently lost or changed jobs. Moreover, many states mandate that particular benefits be part of any health insurance plan.

Discussions of medical care reform in the United States and elsewhere often lead to extreme positions. Advocates at one end believe that the problems with markets in health care are so severe that government control, at least of expenditures, is necessary. The Canadian system – tax-supported, privately provided, but publicly regulated – is held up as an exemplar. The claimed merits are that one insurer eliminates adverse selection, tight supply restrictions manage costs, and tax financing enables everyone to be insured. Of course, in such a system competition between insurers plays no role in promoting efficiency.

At the other extreme are free-market advocates, who believe that market institutions, if guided correctly, would produce a superior outcome. The government should stay out of the insurance business, but implement a risk adjustment system, directly or at arm's length, so that people face efficient prices. Moreover, the government should remove the tax subsidy favoring employer provision of insurance, which would lead to trimmed plan generosity and more cost sharing by employees. Where necessary, the government should give high cost individuals risk-related subsidies that enable them to buy health insurance in the marketplace.

The fundamental difference between the public and private approaches to medical care reform is indicative of the enormous problems in medical care markets and the central role that health plays in our utility. Can risk adjustment work well enough to deter plan manipulation and cream skimming? Without subsidies, would employers provide insurance? If they stopped doing so, how many more people would be uninsured, and how much would their health suffer? These are questions at the heart of health insurance reform.

And beyond the question about organizing the health insurance system, there remain questions of how plans should interact with providers. Should providers be paid by capitation or by fee-for-service, or might there be a happy medium? Will providers respond to a payment schedule by either skimping on patients or driving up costs? Only experience in the future, coupled with a delicately balanced wisdom, will enable us to answer these questions.

Economics does not offer robust conclusions about the virtues and liabilities of markets in second-best situations. Hence, it is not surprising that the debate on who should provide health insurance and how it should best be structured rages on, even among economists. Ultimately, of course, many of the issues cannot be answered on the basis

of first principles, much less the dogma that is too often brought to the debate. They require empirical investigations.

An impressive array of data has been brought to bear one-to-one on central issues in health insurance, but the grand synthesis needed for effective prescription awaits us. Which medical system around the world is best, and what would make it even better? Might the best system for Germany or Japan differ significantly from that for the United States? To understand the attractiveness of alternative health insurance structures, not unlike much of medical care itself, many consequences must be weighed, and many side effects considered. This chapter provided an anatomy to help organize those investigations.

Appendix

This appendix shows the classic treatment of equilibrium with adverse selection and two individuals, from Rothschild and Stiglitz (1976).

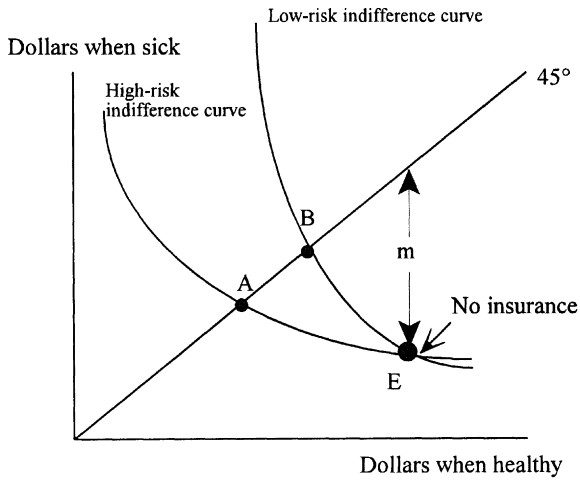
For simplicity, assume that spending when sick, m, is the same for HIGH and LOW, i.e., there is no moral hazard. HIGH is more likely to be sick. Figure A(1) shows the indifference curves for these two people. LOW's indifference curve is steeper than HIGH's, since LOW is not willing to give up as much money when healthy to get a dollar when sick. With no moral hazard, both LOW and HIGH would optimally want full insurance, if charged their fair price for it. Points A and B represent their respective efficient levels of insurance when purchased at actuarially fair rates.

Figure A(2) shows the potential pooling equilibrium. The fair odds line that is shown is the average premium for the two together. At point C, both LOW and HIGH purchase full insurance at this price. But this equilibrium cannot prevail. If an insurer entered the market offering policy D, which has incomplete coverage but a lower premium, he would attract LOW but not HIGH. LOW prefers the policy because he gets the cost savings from not pooling with HIGH, which more than makes up for his loss of full insurance. This is parallel to what happens with the introduction of the basic plan in our numerical analysis, which breaks the pooling equilibrium at moderate.

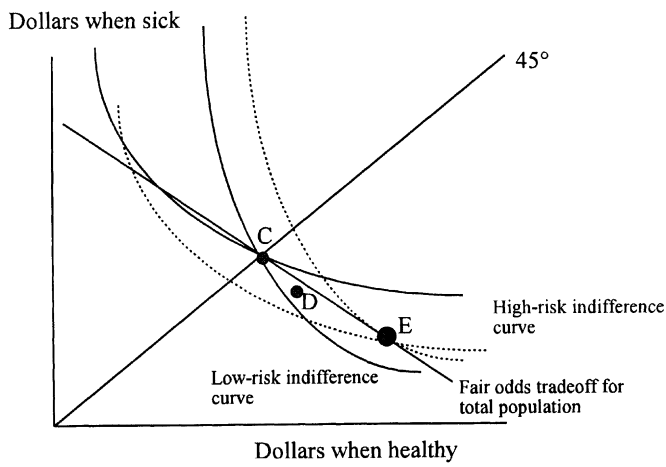
Figure A(3) shows the equilibrium with plan manipulation. HIGH receives full insurance (point A). To separate himself out and thereby reduce his payments, LOW insures incompletely, at point G. Point G makes HIGH just indifferent between staying in the full insurance plan and enrolling in the less generous, but less expensive, policy. Though optimality requires that both groups insure fully, only HIGH does so.

Figure A(4) shows how the separating equilibrium may be broken. We show two fair odds line for the average of HIGH and LOW – one where costs for the two are far apart and one where they are closer together (for simplicity, we show only one indifference curve for HIGH). In the case where HIGH and low have very different costs, the pooled fair odds line will not attract LOW; they do not want to pay the additional amount for more generous coverage because doing so necessitates pooling with HIGH. If the costs are closer together, in contrast, the average fair odds line for the two as a whole

(1) Indifference Curves

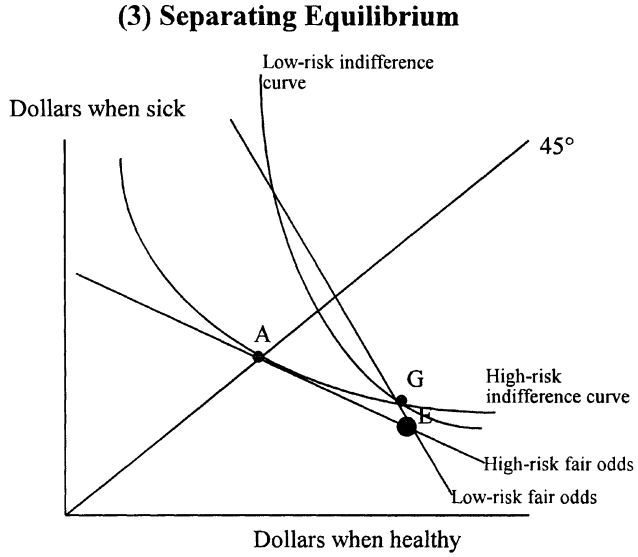


(2) Pooling Equilibrium



Dashed lines are indifference curves through no insurance, point E

Figure A. Adverse selection and plan manipulation.



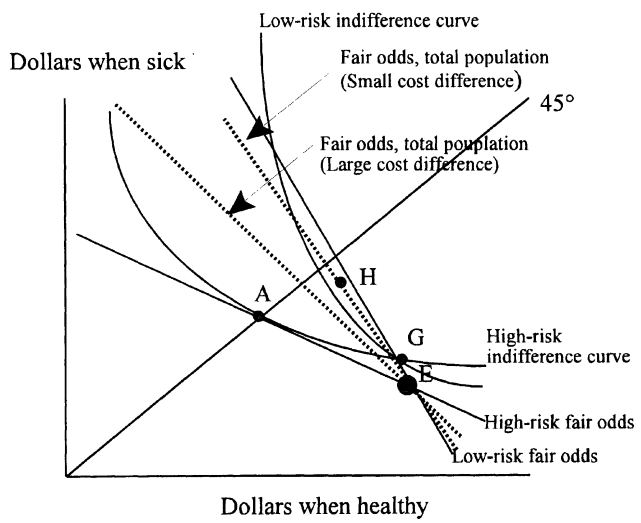


Figure A. (*Continued.*)

will be close to the fair odds line for LOW. Relative to points A and G, there may be a point such as H that will be preferred by LOW to the separating equilibrium. It will also be preferred by HIGH, who benefits from pooling with the healthier group in the population. It will thus undermine the separating equilibrium. With no stable pooling equilibrium and no stable separating equilibrium, the market will not reach an equilibrium.

References

Akerlof, G. (1970), "The market for 'Lemons': qualitative uncertainty and the market mechanism", Quarterly Journal of Economics 74:488–500.

Altman, D., D.M. Cutler and R.J. Zeckhauser (1998), "Adverse selection and adverse retention", American Economic Review 88(2):122–126.

Arrow, K. (1963), "Uncertainty and the welfare economics of medical care", American Economic Review 53(5):941–973.

Arrow, K. (1965), Aspects of the Theory of Risk Bearing (Yrjo Jahnssonin Saatio, Helsinki).

Arrow, K. (1985), "The economics of agency", in: J. Pratt and R. Zeckhauser, eds., Principals and Agents: The Structure of Business (Harvard Business School Press, Cambridge, MA) 37–51.

Baker, L.C., and K.S. Corts (1995), "The effects of HMOs on conventional insurance premiums: theory and evidence", NBER Working Paper No. 5356.

Baumgardner, J. (1991), "The interaction between forms of insurance contract and types of technical change in medical care", Rand Journal of Economics 22(1):36–53.

Beck, R.G. (1974), "The effects of co-payment on the poor", Journal of Human Resources 9(1):129–142.

Berk, M.L., and A.C. Monheit (1992), "The concentration of health expenditures: an update", Health Affairs 11(4):145–149.

Bhattacharya, J., W.B. Vogt, A. Yoshikawa and T. Nakahara (1996), "The utilization of outpatient medical services in Japan", Journal of Human Resources 31(2):450–476.

Bice, T.W. (1975), "Risk vulnerability and enrollment in a prepaid group practice", Medical Care 13(8):698–703.

Blomqvist, A.G. (1997), "Optimal non-linear health insurance", Journal of Health Economics 16(3):303–321.

Brown, R.S., et al. (1993), "The Medicare risk program for HMOs – Final summary report on findings from the evaluation", Final Report under HCFA Contract No. 500-88-0006 (Mathematica Policy Research, Inc., Princeton, NJ).

Buchanan, J.L., E.B. Keeler, J.E. Rolph and M.R. Holmer (1991), "Simulating health expenditures under alternative insurance plans", Management Science 37(9):1067–1089.

Buchmueller, T.C., and P.J. Feldstein (1997), "The effect of price on switching among health plans", Journal of Health Economics 16(2):231–247.

Cardon, J., and I. Hendel (1996), "Asymmetric information in health care and health insurance markets: evidence from the National Medical Expenditure Survey", mimeo.

Carroll, N.V., and W.G. Erwin (1987), "Patient shifting as a response to Medicare prospective payment", Medical Care 25(12):1161–1167.

Cave, J. (1985), "Subsidy equilibrium and multiple-option insurance markets", in: R. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research. Biased Selection in Health Care Markets, Vol. 6 (JAI Press, Greenwich, CT) 27–45.

Cherkin, D.C., L. Grothaus and E.H. Wagner (1989), "The effect of office visit copayments on utilization in a health maintenance organization", Medical Care 27(7):669–679.

Cochrane, J. (1995), "Time consistent health insurance", Journal of Political Economy 103(3):445–73.

Colle, A.D., and M. Grossman (1978), "Determinants of pediatric care utilization", Journal of Human Resources 13(Suppl.):115–153.
Conrad, D.A., D. Grembowski and P. Milgrom (1985), "Adverse selection within dental insurance markets", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 171–190.
Congressional Budget Office (1992), "The effects of managed care on use and costs of health services", mimeo.
Cutler, D.M. (1991), "Estimating the effect of reimbursement policy on medical outcomes", Doctoral dissertation (Massachusetts Institute of Technology, MA).
Cutler, D.M. (1995), "The incidence of adverse medical outcomes under prospective payment", Econometrica 63(1):29–50.
Cutler, D.M. (1996), "Why don't markets insure long-term risk?", mimeo.
Cutler, D.M., M. McClellan and J.P. Newhouse (1998), "What does managed care do?", mimeo.
Cutler, D.M., and S.J. Reber (1998), "Paying for health insurance: the tradeoff between competition and adverse selection", Quarterly Journal of Economics 113(2):433–466.
Cutler, D.M., and R.J. Zeckhauser (1998), "Adverse selection in health insurance", in: A. Garber, ed., Frontiers in Health Policy Research, Vol. 1 (MIT Press, Cambridge, MA) 1–31.
Davis, K., and L.B. Russell (1972), "The substitution of hospital outpatient care for inpatient care", Review of Economics and Statistics 54(2):109–120.
De Meza, D. (1983), "Health insurance and the demand for medical care", Journal of Health Economics 2(1):47–54.
DesHarnais, S.I., J. Chesney and S. Fleming (1988), "Trends and regional variations in hospital utilization and quality during the first two years of the prospective payment system", Inquiry 25:374–382.
DesHarnais, S.I., R. Wroblewski and D. Schumacher (1990), "How the Medicare prospective payment system affects psychiatric patients treated in short-term general hospitals", Inquiry 27:382–388.
Diamond, P. (1998), "Optimal income taxation: an example with a U-shaped pattern of optimal marginal tax rates", American Economic Review 88(1):83–95.
Dowd, B., and R. Feldman (1985), "Biased selection in twin cities health plans", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 253–271.
Eggers, P.W. (1980), "Risk differential between Medicare beneficiaries enrolled and not enrolled in an HMO", Health Care Financing Review 1:91–99.
Eichner, M.J. (1998), "Incentives, price expectations and medical expenditures: an analysis of claims under employer-provided health insurance", mimeo.
Eichner, M.J., M. McClellan and D. Wise (1998), "Insurance or self-insurance?: Variation, persistence, and individual health accounts", in: D. Wise, ed., Inquiries in the Economics of Aging (University of Chicago Press, Chicago, IL) 19–45.
Ellis, R.P. (1985), "The effect of prior-year health expenditures on health coverage plan choice", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 127–147.
Ellis, R.P. (1989), "Employee choice of health insurance", Review of Economics and Statistics 71(2):215–223.
Ellis, R.P. (1998), "Creaming, skimping and dumping: provider competition on the intensive and extensive margins", Journal of Health Economics 17(5):537–555.
Ellis, R.P., and T.G. McGuire (1986), "Provider behavior under prospective reimbursement: cost sharing and supply", Journal of Health Economics 5(2):129–152.
Ellis, R.P., and T.G. McGuire (1996), "Hospital response to prospective payment: moral hazard, selection, and practice-style effects", Journal of Health Economics 15:257–277.
Eze, P., and B. Wolfe (1993), "Is dumping socially inefficient? An analysis of the effect of Medicare's prospective payment system on the utilization of Veterans Affairs inpatient services", Journal of Public Economics 52:329–344.

Farley, P.J., and A.C. Monheit (1985), "Selectivity in the demand for health insurance and health care", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, vol. 6 (JAI Press, Greenwich, CT) 231–252.

Feder, J., J. Hadley and S. Zuckerman (1987), "How did Medicare's prospective payment system affect hospitals?", New England Journal of Medicine 317(14):867–873.

Feldman, R., and D. Dowd (1991), "Must adverse selection cause premium spirals?", Journal of Health Economics 10(3):350–357.

Feldman, R., and B. Dowd (1993), "The effectiveness of managed competition in reducing the costs of health insurance", in: R.B. Helms, ed., Health Policy Reform: Competition and Controls (AEI Press, Washington, DC) 176–217.

Feldman, R., M. Finch, B. Dowd and S. Cassou (1989), "The demand for employment-based health insurance plans", Journal of Human Resources 24(1):117–142.

Feldstein, M.S. (1970), "The rising price of physicians' services", Review of Economics and Statistics 52(2):121–133.

Feldstein, M.S. (1971), "Hospital cost inflation: a study of nonprofit price dynamics", American Economic Review 60:853–872.

Feldstein, M.S., and B. Friedman (1977), "Tax subsidies, the rational demand for insurance and the health care crisis", Journal of Public Economics 7(2):155–178.

Feldstein, P.J. (1964), "General report", Report of the Commission on the Cost of Medical Care, Part 1 (American Medical Association, Chicago).

Fisher, C.R. (1992), "Hospital and Medicare financial performance under PPS, 1985–90", Health Care Financing Review 14(1):171–183.

Fitzgerald, J.F., L.F. Fagan, W.M. Tierney and R.S. Dittus (1987), "Changing patterns of hip fracture care before and after implementation of the prospective payment system", Journal of the American Medical Association 258(2):218–221.

Fitzgerald, J.F., P.S. Moore and R.S. Dittus (1988), "The care of elderly patients with hip fracture: changes since implementation of the prospective payment system", New England Journal of Medicine 319(21):1392–1397.

Folland, S., and R. Kleiman (1990), "The effect of prospective payment under DRGs on the market value of hospitals", Quarterly Review of Economics and Business 30(2):50–68.

Frank, R.G., and J.R. Lave (1986), "The effect of benefit design on the length of stay of Medicaid psychiatric patients", Journal of Human Resources 21(3):321–337.

Frank, R.G., and J.R. Lave (1989), "A comparison of hospital responses to reimbursement policies for Medicaid psychiatric patients", RAND Journal of Economics 20(4):588–600.

Frank, R.G., and T. McGuire (1998), "Economic functions of carve-outs", American Journal of Managed Care 4(SP):SP31–SP39.

Frank, R.G., T.G. McGuire and J.P. Newhouse (1995), "Risk contracts in managed mental health care", Health Affairs 14(3):50–64.

Frank, R.G., J. Glazer and T.G. McGuire (1998), "Measuring adverse selection in managed health care", NBER Working Paper no. 6825, December.

Fuchs, V.R., and M.J. Kramer (1972), "Determinants of expenditures for physicians' services in the United States, 1948–68", National Bureau of Economic Research Occasional Paper Series, No. 117.

Gaumer, G.L., E.L. Poggio, C.G. Coelen, C.S. Sennett and R.J. Schmitz (1989), "Effects of state prospective reimbursement programs on hospital mortality", Medical Care 27(7):724–736.

Gerety, M.B., V. Soderholm-Difatte and C.H. Winograd (1989), "Impact of prospective payment and discharge location on the outcome of hip fracture", Journal of General Internal Medicine 4(5):388–391.

Glied, S. (2000), "Managed care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 13.

Goldman, F., and M. Grossman (1978), "The demand for pediatric care: an hedonic approach", Journal of Political Economy 86(2):259–280.

Griffith, M.J., N. Baloff and E.L. Spitznagel (1984), "Utilization patterns of health maintenance organization disenrollees", Medical Care 22(9):827–834.

Guterman, S., S.H. Altman and D.A. Young (1990), "Hospitals' financial performance in the first five years of PPS", Health Affairs 9(1):125–134.

Guterman, S., and A. Dobson (1986), "Impact of the Medicare prospective payment system for hospitals", Health Care Financing Review 7(3):97–114.

Hadley, J., S. Zuckerman and J. Feder (1989), "Profits and fiscal pressure in the prospective payment system: their impacts on hospitals", Inquiry 26:354–365.

Hodgkin, D., and T.G. McGuire (1994), "Payment levels and hospital response to prospective payment", Journal of Health Economics 13:1–29.

Hurley, J. (2000), "An overview of normative economics of the health sector", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 2.

Jackson-Beeck, M., and J.H. Kleinman (1983), "Evidence for self-selection among health maintenance organization enrollees", Journal of the American Medical Association 250(20):2826–2829.

Juba, D.A., J.R. Lave and J. Shaddy (1980), "An analysis of the choice of health benefits plans", Inquiry 17:62–71.

Kahn, K.L., et al. (1990) (series), "The effects of the DRG-based prospective payment system on quality of care for hospitalized Medicare patients", Journal of the American Medical Association 264(15):1953–1994 (eight articles).

Keeler, E.B., J.P. Newhouse and C.E. Phelps (1977), "Deductibles and demand: a theory of the consomer facing a variable price schedule under uncertainty", Econometrica 45:641–655.

Keeler, E.B., G. Carter and J.P. Newhouse (1998), "A model of the impact of reimbursement schemes on health plan choice", Journal of Health Economics 17(3):297–320.

Kotowitz, Y. (1987), "Moral hazard", in: New Palgrave Dictionary of Economics.

Langwell, K.M., and J.P. Hadley (1989), "Evaluation of the Medicare competition demonstrations", Health Care Financing Review 11(2):65–80.

Lave, J.R., R.G. Frank, C. Taube, H. Goldman and A. Rupp (1988), "The early effects of Medicare's prospective payment system on psychiatry", Inquiry 25:354–363.

Long, S.H., R.F. Settle and C.W. Wrightson (1988), "Employee premiums, availability of alternative plans, and HMO disenrollment", Medical Care 26(10):927–938.

Luft, H.S., J.B. Trauner and S.C. Maerki (1985), "Adverse selection in a large, multiple-option health benefits program: a case study of the California Public Employees' Retirement System", in: R.M. Scheffler and L.F. Rossiter, eds., Advances in Health Economics and Health Services Research, Vol. 6 (JAI Press, Greenwich, CT) 197–229.

Ma, C.A., and T. McGuire (1997), "Optimal health insurance and provider payment", American Economic Review 87(4):685–704.

Maibach, E., K. Dusenbury, P. Zupp et al. (1998), "Marketing HMOs to Medicare Beneficiaries: An Analysis of Four Medicare Markets" (Kaiser Family Foundation, Menlo Park, CA).

Manning, W.G., and M.S. Marquis (1996), "Health insurance: the tradeoff between risk pooling and moral hazard", Journal of Health Economics 15(5):609–639.

Marquis, M.S. (1992), "Adverse selection with a multiple choice among health insurance plans: a simulation analysis", Journal of Health Economics 11(2):129–151.

Marquis, M.S., and C.E. Phelps (1987), "Price elasticity and adverse selection in the demand for supplemental health insurance", Economic Inquiry 25(2):299–313.

McAvinchey, I.D., and A. Yannopoulos (1993), "Elasticity estimates from a dynamic model of interrelated demands for private and public acute health care", Journal of Health Economics 12(2):171–186.

McGuire, T.G. (1981), "Price and membership in a prepaid group medical practice", Medical Care 19(2):172–183.

McGuire, T.G. (2000), "Physician agency", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 9.

Menke, T. (1990), "Impacts of PPS on Medicare Part B expenditures and utilization for hospital episodes of care", Inquiry 27(2):114–126.

Merrill, J., C. Jackson and J. Reuter (1985), "Factors that affect the HMO enrollment decision: a tale of two cities", Inquiry 22(4):388–395.

Miller, R.H., and H.S. Luft (1997), "Does managed care lead to better or worse quality of care?", Health Affairs 16(5):7–25.

Mirrlees, J.A. (1971), "An exploration in the theory of optimum income taxation", Review of Economic Studies 38:175–208.

Morrisey, M.A., F.A. Sloan and J. Valvona (1988), "Medicare prospective payment and posthospital transfers to subacute care", Medical Care 26(7):685–698.

Newhouse, J.P. (1989), "Do unprofitable patients face access problems?", Health Care Financing Review 11(2):33–42.

Newhouse, J.P. (1996), "Reimbursing health plans and health providers: efficiency in production versus selection", Journal of Economic Literature 34(3):1236–1263.

Newhouse, J.P., and the Insurance Experiment Group (1993), Free for All? Lessons from the RAND Health Insurance Experiment (Harvard University Press, Cambridge, MA).

Newhouse, J.P., and D.J. Byrne (1988), "Did Medicare's prospective payment system cause length of stay to fall?", Journal of Health Economics 7(4):413–416.

Newhouse, J.P., and C.E. Phelps (1974), "Price and income elasticities for medical care services", The Economics of Health and Medical Care (John Wiley & Sons, New York), ch. 9, 140–161.

Newhouse, J.P., and C.E. Phelps (1976), "New estimates of price and income elasticities of medical care services", The Role of Health Insurance in the Health Services Sector (National Bureau of Economic Research, New York), Chapter 7, 261–313.

Nichols, A., and R. Zeckhauser (1982), "Targeting transfers through restrictions on recipients", American Economic Review 72(2):372–377.

Palmer, R.M., R.M. Saywell Jr., T.W. Zollinger, B.K. Erner, A.D. LaBov, D.A. Freund, J.E. Garber, G.W. Misamore and F.B. Throop (1989), "The impact of the prospective payment system on the treatment of hip fractures in the elderly", Archives of Internal Medicine 149(10):2237–2241.

Pauly, M. (1968), "The economics of moral hazard: comment", American Economic Review 58:531–536.

Pauly, M. (1974), "Overinsurance and public provision of insurance: the roles of moral hazard and adverse selection", Quarterly Journal of Economics 88(1):44–54.

Pauly, M. (1986), "Taxation, health insurance and market failure", Journal of Economic Literature 24(2):629–675.

Pauly, M., H. Kunreuther and R. Hirth (1995), "Guaranteed renewability in insurance", Journal of Risk & Uncertainty 10(2):143–156.

Pauly, M., and S. Ramsey (1998), "Would you like suspenders to go with that belt? An analysis of optimal combinations of cost sharing and managed care", mimeo.

Phelps, C.E. (1973), "The demand for health insurance: a theoretical and empirical investigation", RAND Research Paper Series, No. R-1054-OEO.

Phelps, C.E., and J.P. Newhouse (1972a), "Effect of coinsurance: a multivariate analysis", Social Security Bulletin 20–28.

Phelps, C.E., and J.P. Newhouse (1972b), "Effects of coinsurance of demand for physician services", RAND Research Paper Series, No. R-976-OEO.

Phelps, C.E., and J.P. Newhouse (1974), "Coinsurance, the price of time, and the demand for medical services", Review of Economics and Statistics 56(3):334–342.

Plato, The Republic.

Price, J.R., and J.W. Mays (1985), "Biased selection in the Federal Employees Health Benefits Program", Inquiry 22(1):67–77.

Ramsey, S.D., and M. Pauly (1997), "Structural incentives and adoption of medical technologies in HMO and fee-for-service health insurance plans", Inquiry 34(3):228–236.

Ray, W.A., M.R. Griffin and D.K. Baugh (1990), "Mortality following hip fracture before and after implementation of the prospective payment system", Archives of Internal Medicine 150(10):2109–2114.

Rodgers, J., and K.E. Smith (1996), "Is there biased selection in Medicare HMOs?", Health Policy Economics Group Report (Price Waterhouse LLP, Washington, DC).

Roos, N.P., E. Shapiro and R. Tate (1989), "Does a small minority of elderly account for a majority of health care expenditures? A sixteen-year perspective", Milbank Quarterly 67(3-4):347–369.

Rosenthal, G. (1970), "Price elasticity of demand for short-term general hospital services", in: H.E. Klarman, ed., Empirical Studies in Health Economics (Johns Hopkins Press, Baltimore, MD) 101–117.

Rosett, R.N., and L. Huang (1973), "The effect of health insurance on the demand for medical care", Journal of Political Economy 81(March/April):281–305.

Rothschild, M., and J.E. Stiglitz (1976), "Equilibrium in competitive insurance markets: an essay on the economics of imperfect information", Quarterly Journal of Economics 90(4):630–649.

Russell, L.B., and C.L. Manning (1989), "The effect of prospective payment on Medicare expenditures", New England Journal of Medicine 320(7):439–444.

Sager, M.A., D.V. Easterling, D.A. Kindig and O.W. Anderson (1989), "Changes in the location of death after passage of Medicare's prospective payment system", New England Journal of Medicine 320(7):433–439.

Scitovsky, A.A., and N. McCall (1977), "Coinsurance and the demand for physician services: four years later", Social Security Bulletin 19–27.

Scitovsky, A.A., N. McCall and L. Benham (1978), "Factors affecting the choice between two prepaid plans", Medical Care 16(8):660–675.

Scitovsky, A.A., and N.M. Snyder (1972), "Effect of coinsurance on use of physician services", Social Security Bulletin 3–19.

Shaw, G.B. (1911), The Doctors Dilemma.

Sheingold, S.H. (1989), "The first three years of PPS: impact on Medicare costs", Health Affairs 8(3):191–204.

Sheingold, S.H., and T. Buchberger (1986), "Implications of Medicare's prospective payment system for the provision of uncompensated hospital care", Inquiry 23(4):371–381.

Sloan, F.A., M.A. Morrisey and J. Valvona (1988), "Medicare prospective payment and the use of medical technologies in hospitals", Medical Care 26(9):837–850.

Smith, A. (1776), The Wealth of Nations.

Spence, M., and R. Zeckhauser (1971), "Insurance, information, and individual action", American Economic Review 61(2):380–387.

Staiger, D., and G.L. Gaumer (1995), "Price regulation and patient mortality in hospitals", mimeo.

van de Ven, W.P.M.M., and R.P. Ellis (2000), "Risk adjustment in competitive health plan markets", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 14.

van de Ven, W.P.M.M., and R.C.J.A. van Vliet (1995), "Consumer surplus and adverse selection in competitive health insurance markets: an empirical study", Journal of Health Economics 14(2):149–169.

Wagstaff, A., and E.K.A. van Doorslaer (2000), "Equity in health care finance and delivery", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 34.

Welch, W.P. (1989), "Restructuring the Federal Employees Health Benefits Program: the private sector option", Inquiry 26(3):321–334.

Weissman, J., and A. Epstein (1994), Falling Through the Safety Net: Insurance and Access to Medical Care (Johns Hopkins University Press, Baltimore, MD).

Wholey, D., R. Feldman and J.B. Christianson (1995), "The effect of market structure on HMO premiums", Journal of Health Economics 14(1):81–105.

Williams, A., and R. Cookson (2000), "Equity in health", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 35.

Willke, R.J., W.S. Custer, J.W. Moser and R.A. Musacchio (1991), "Collaborative production and resource allocation: the consequences of prospective payment for hospital care", Quarterly Review of Economics and Business 31(1):28–47.

Wilson, C. (1980), "The nature of equilibrium in markets with adverse selection", Bell Journal of Economics 11(1):108–130.

Wrightson, W., J. Genuardi and S. Stephens (1987), "Demographic and utilization characteristics of HMO disenrollees", GHAA Journal 23–42.

Zeckhauser, R. (1970), "Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives", Journal of Economic Theory 2(1):10–26.

Zeckhauser, R. (1974), "Risk spreading and distribution", in: H.M. Hochman and G.E. Peterson, eds., Redistribution Through Public Choice (Columbia University Press, New York) 206–228.

Zweifel, P., and W.G. Manning (2000), "Moral hazard and consumer incentives in health care", in: A.J. Culyer and J.P. Newhouse, eds., Handbook of Health Economics (Elsevier, Amsterdam) Chapter 8.