Oxford Handbooks Online

Insurance and the Demand for Medical Care

Mark V. Pauly

The Oxford Handbook of Health Economics
Edited by Sherry Glied and Peter C. Smith

Print Publication Date: Apr 2011
Subject: Economics and Finance, Health, Education, and Welfare, Public Economics and Policy
Online Publication Date: Sep 2012 DOI: 10.1093/oxfordhb/9780199238828.013.0016

Abstract and Keywords

Insurance coverage affects the use and cost of medical care, and so potentially can play a role in assuring that spending comes closer to the optimum. This article describes the implications of third party financing, whether public or private. The key issue is that—in the absence of direct user payment for services—there is an incentive for inefficient moral hazard, or excess use of services. This article uses the voluntary insurance purchasing model to frame the discussion of demand effects because that is the model used extensively in the literature. It later raises the alternative social goals model and also uses this to interpret insurance effects on demand. This discussion implies that consumers will demand the most generous insurance coverage against types of care or types of illnesses for which demand responsiveness is low, the probability of illness is low, and the cost of treatment is high.

Keywords: insurance, medical care, moral hazard, social goals model, demand, illnesses

16.1 Introduction

INSURANCE coverage of the cost of or spending on medical goods and services will often change the quantities, qualities, types, or prices of the care patients receive, relative to what would have happened had they not had insurance. Among those with some insurance, differences in the form or extent of coverage can also affect use. Research directed at understanding when and why health insurance affects the demand for medical care (and therefore the price or quantity of what is demanded) is extensive. If one views this research (as we shall suggest below) as tracing out the consumer or market demand

function for care, this too has been one of the most intensively investigated demand curves in all of applied microeconomics. This chapter will discuss what is known about this phenomenon and how economic theory interprets what is known.

There are two different policy questions which information on the relationship between insurance and demand may help to answer. The one that has been the primary focus of the bulk of such research has concerned the effect of insurance on the demand for *care* (called moral hazard in the insurance literature). A second question takes the analysis a step further and explores the impact of any insurance-care demand connection on the voluntary demand for insurance. The first line of inquiry is relevant in countries with both voluntary and mandatory health insurance, since it deals with the impact of variations in insurance design and coverage on the use of care. The second issue has been more extensively treated in settings in which insurance coverage is (to some extent) voluntary, but (as I will argue) it is also relevant to the choice of optimal insurance design. (This health economics research is not relevant in countries where (p. 355) public policy specifies that insurance must fully cover all services with no cost sharing or opportunities for out-of-pocket payment for non-covered services.)

Models of the demand for health insurance almost always incorporate analysis of the effect of insurance on demand for care, called “moral hazard” in the insurance literature.[1] In many (though, as we shall see, by no means all) cases, a larger potential or actual impact of coverage on demand (compared to a small or zero impact) reduces the buyer's demand for insurance. The possibility of moral hazard in which insurance affects behavior which in turn alters expected losses or claims poses a potential trade-off between moral hazard and risk protection in private or public insurance demand.

Strong demand effects potentially have negative effects on welfare when insurance is voluntary and the person's use of medical care and health only affects the person's utility. But for at least some populations in developed countries, insurance is either heavily subsidized or mandated primarily because of a desire to change the demand for care (as well as to provide financial protection). In this case, the policy question and social goal (and potential efficiency improvement) involves creating *more* moral hazard than would otherwise exist, and harnessing it in ways that achieve social goals.

In what follows I will initially use the voluntary insurance purchasing model to frame the discussion of demand effects because that is the model used extensively in the literature. But I will (much) later raise the alternative social goals model (though it is much less definitively and rigorously specified than the voluntary purchase model), and also use this to interpret insurance effects on demand.

16.2 Introduction to Moral Hazard in Insurance Economics

Voluntary insurance for any kind of risk works best and most efficiently in cases when there is a chance that a loss-producing event may occur, but when both the amount of the loss and the chance that it will happen are fixed, regardless of the behavior of the insurance purchaser. Medical expense insurance often does not quite fit this case. There is no doubt that illnesses occur to some extent randomly, and that some medical spending may be desirable to pay for treatment of the illness. But both the chances of getting sick and the amount of medical care and medical spending associated with the illness are to some extent under the control of the consumer, so the consumer (on the demand side) may choose to change both the illness probability and the amount of medical spending to different levels with and without insurance coverage. On the supply side, the possibility of variation in treatment means that provider recommendations for treatment (and cost) may also vary with the presence of insurance and the way in which providers are paid. When such moral (p. 356) hazard is possible, the consumer faces a trade-off: more protection against risk may cause the consumer to behave differently, in ways that increase money losses or make given losses more likely to happen, and which in turn raise the cost of insurance (Zeckhauser 1970). There are therefore two main questions about voluntary insurance design for individual consumers—both positive as describing the insurance people will choose and normative as defining the insurance they should choose for efficiency: (1) under a given form of insurance, what risk protection of what type should be sacrificed to control moral hazard? (2) How might insurance be configured best in view of the trade-off between risk protection and moral hazard? For the most part, analysis of the effect of insurance on patient demand has taken that demand as determined by the patient's illness state, preferences, and insurance. I will consider possible effects on supplier recommendations or advice for treatment (induced demand), although the effect of insurance on supply is really a different (and less extensively investigated) topic. The classic treatment of that subject is by Ma and McGuire 1997. I will not deal extensively with that subject here.

16.3 The (Probably Infeasible) Benchmark Indemnity Model

Is there a setting in which insurance would have no effects on demand? One might imagine that there could be health insurance which would make dollar payments of a predetermined amount, depending on the person's health state. Conditional on the occurrence of a given illness, the consumer (with advice from the physician) will choose a level of use of medical services and products and an associated level of spending so that consumption of all types of medical care is set at the point where the marginal “health and welfare” benefit of care equals the marginal cost of that care in a competitive market. The level of spending thus determined will then be the dollar amount the consumer will want the insurance to pay if the illness strikes.

To make this idea concrete, first think of a patient-consumer who knows everything that is known about the effectiveness of treatment of various amounts, types, and costs for a given illness, and about the effectiveness of treatment of different illnesses of different degrees of severity. Suppose there is just one illness of a uniform severity that might affect this person with probability p. Without insurance, this person would choose how much to spend ($S^\star$) on medical care by comparing the marginal benefits from additional spending with the cost of that spending for each type of medical care; call that package of care $X^\star$. Finally, suppose the person could buy insurance that would make a payment when the illness hits and not otherwise, and whose premium is equal to the expected value of claims payments (actuarially fair). When premiums are actuarially fair, insurance will be demanded by all risk-averse people, and they will demand the same level of coverage regardless of the strength of risk aversion. But even if premiums are fair variations (p. 357) in risk aversion will affect insurance demand when moral hazard is present. What payment (or benefit level) would the person most prefer?

It is easy to see that it is approximately $S^\star$. If the person gets a check for $S^\star$, he will then want to choose $X^\star$, the quantity of care at which marginal benefit equals marginal cost. Setting aside (for the present) some usually small income effects, that will be the level of spending at which marginal benefit originally (without insurance) equaled marginal cost. If we ignore insurance administrative cost, the premium will then be p dollars per dollar of benefit to be paid; at that unit premium a risk-averse person will choose the benefit level B as equal to $S^\star$, and therefore choose to pay a total premium $P = pS^\star$. Not only will that coverage eliminate financial risk (in the sense that the amount the consumer will have available to spend on other consumption will be the same whether ill or not), but the quantity of care the consumer will want to have will be the quantity whose cost can be covered by the insurance payment. That is, at $B = S^*$, the consumer will want to consume just $S^\star$ worth of care. Consuming more will be unattractive because the expected benefit will fall short of the additional cost, and consuming less (while pocketing some of the

insurance payment) will also not happen because the consumer will value the care more than the money.

As an alternative to paying the dollar amount $S^\star$, insurers might offer a “managed care” insurance contract, which takes the form of promising a certain amount and type of care conditional on the occurrence of each illness. It is clear that the optimal managed care benefit in this simple model is care that costs $\$S^\star$ and is the same as $X^\star$. This makes the point that in a world of perfect information there is no intrinsic advantage to managed care, since it produces a final result that is identical to that of optimal indemnity insurance.

16.4 Necessary Conditions for Indemnity Insurance to be Feasible and Optimal

In this simple model there was only one illness and one level of severity, so it might be plausible to assume that the insurer could tell if the insured was sick or not. If there are multiple illnesses and multiple levels of severity for a given illness, the theoretically ideal indemnity insurance would be one that made payments conditional on the existence, type and severity of illness, at the level of benefit $S^\star$ for that illness, severity, and type. While I believe that there is more scope for the use of pure indemnity contracts than now exists, there are limitations to them. The most obvious one is that the insurer may not be able to know precisely what illness and severity level occurs. Suppose at some severity level L the person would want $S^\star$ spent on medical care, and at some more severe level L the person would want to spend S. However, in the absence of any additional out-of-pocket cost, the person may prefer S even in the state L because the additional spending provides positive marginal benefit.

(p. 358) It is now easy to see what the problem is. The consumer will claim to be at severity level L' regardless, and the average or expected value of care with insurance will therefore be higher than without insurance: moral hazard will occur. Moral hazard is one of the distinguishing features of health insurance. It is also easy to see that moral hazard leads to consumption of care worth less than its benefit: in state L with insurance the person consumes $S' - S^\star$ additional care, but we know that this additional care is not worth what it costs. Yet it will cause the premium to rise from $p(L)\ S^\star + p(L')S'$ to $(p(L) + p(L'))S'$, a larger number (where $p(L)$ is the probability of getting the illness at the severity level L and $p(L')$ is the probability of getting it at the higher severity level L).

More generally, moral hazard arises when the insurer cannot tell exactly how sick the person is, and therefore does not know which of many marginal benefit curves is applicable to the situation in question (Pauly 2008). The mirror image of this statement is also illuminating: moral hazard need not occur if the insurer can tell what the person's illness condition is. There has been some discussion of cases in which the insurer is assumed to have this knowledge, and suggesting that managed care rules or value-based cost sharing should then be applied (Chernew, Rosen, and Fendrick 2007). But note that in the perfect information case, managed care rules would say that the person should receive $S^\star$ units of care, while value based insurance should pay the full cost of the $S^\star$ amount of care. That is, there should be no positive cost sharing at all: insurance should pay in full for the amount of care appropriate to the patient's condition, and not pay for any other course of treatment.

In the more realistic case when the insurer cannot know precisely the state of health, the optimal benefit payment (or managed care treatment package) will be a compromise between the benefit payment if L occurs and the benefit payment if L' occurs. If S and S' would represent the optimal spending levels corresponding to each state, in the world

with moral hazard the insurance might pay something between S and S', with the result that in the truly more sick state S' there will be positive patient cost sharing, and consequently exposure to financial risk. The greater the degree of risk aversion, the higher the payment, other things equal. If the insurance requires that the full amount of any benefit be spent on medical care, there will also be over-use of care (relative to that at which marginal benefit equals marginal cost) in the less severe illness state. Alternatively, the managed care policy might specify a treatment package between the two ideal levels; there will be no financial risk, but there will be the risk of under-provision of care in the S' state, and possible over-provision in the S state if patients always adhere to the specified treatment package. Financial risk will be replaced by health outcome risk.

16.5 Income Effects with Moral Hazard

The purpose of insurance is to change the household's level of wealth in different states of the world, compared to a situation with no insurance. Wealth is increased in “high loss” states where benefits exceed premiums, and reduced in “low loss” states where premiums exceed benefits. If insurance is actuarially fair, expected wealth is left unchanged, and if (p. 359) the buyer is risk averse, expected utility is increased. However, it is likely that the amount of medical spending at which the household's marginal willingness to pay (amount of other consumption sacrificed for medical care) will be larger in a high loss state with insurance benefit payment received than without insurance because the benefit payment will affect the level of wealth or income. In the indemnity case, although much of the benefit payment will be used to pay for the medical services that would have been bought in the absence of insurance, some of what would have been a reduction in consumption of other goods without insurance may also be diverted to medical services if the higher wealth raises the person's willingness to pay for additional wealth. That is, rather than return to the level of consumption spending that would have occurred in the absence of a loss, the person may choose to cut into consumption to some extent. There may well then be positive income or wealth effects on spending in the high loss case; by the same argument, there should be negative effects on medical spending in the low loss case.

Will this change in the pattern of use with insurance mean that overall medical spending will rise or fall? There are two potential channels for income effects. One is from the premium. If, on the one hand, the insurance is subsidized, that will increase expected wealth compared to a world with no insurance and no subsidy. On the other hand, if the insurance premium rises, either because of higher loading or because of higher medical care prices, that will reduce wealth, other things equal. Of course the risk-averse person gets a gain in expected utility from insurance even if the premium has a loading.

The other channel concerns the relationship between the premium and the insurance payout in the particular health state that occurs. If the benefit payment exceeds the premium, wealth is increased in that state. If the payment is below the premium, wealth is reduced. The net effect of coverage on use obviously depends on whether the income effects in the high loss states cancel out the income effects in the low loss states. Generally they will not cancel out; rather, the net income effect on average spending per insured person (compared to average spending per uninsured person) will be positive as Nyman 1999 following de Meza 1983 has shown most clearly. If there is a uniform income elasticity of demand and if there is positive health care spending in all states, and if the premium is close to fair, it is easy to see that the percentage change in wealth from insurance in high loss states, weighted by the probability of being in those states, will approximately equal the weighted percentage change in wealth in low loss states (or fall a little short if there is insurance loading). For example suppose that the probability of the high loss state is 0.1, the person's wealth is 100, the uninsured expense in the high loss state is 50 and in the low loss state is 5. The premium is assumed to be actuarially fair at 9.5. Thus wealth is increased in the high loss state (compared to no insurance) by 40.5, and is reduced in the low loss state by 4.5. The percentage change in wealth (relative to the base of 100) is thus ten times greater (in the opposite direction) in the high loss state compared to the low loss state. But while the percentage changes in wealth are approximately offsetting, the amounts of spending are different: a given percentage change in spending is obviously much greater in a high spending state than in a low spending state. So the overall impact of changes in *ex post* wealth on the quantity of care demanded and on spending is highly likely to be positive.

(p. 360) The possibility of a disproportionately larger change in spending in the case of a rare high cost illness is plausible. One reason is the household's wealth acts as a limit on total spending (actually wealth minus minimum subsistence); in the example spending cannot exceed 100. But the household clearly would not violate its budget constraint if it bought insurance that paid 200 in the high loss state, since the premium for such insurance (24.5) fits within the wealth constraints of either state.

There is thus a modest difference between the theory of insurance as applied to risks that only reduce money wealth (where the monetary value of the reduction in well-being cannot exceed uninsured wealth) compared to insurance that pays directly for higher consumer spending when adverse events occur. But this alone is not a major distinguishing feature between health insurance and other insurance markets thought to function in a more satisfactory way.

In the first case, suppose a person has wealth of $\$W$ and a risk of a loss with probability p. It seems obvious that the worst thing that can happen to that person is that his wealth is driven to zero. It therefore follows that the person will never demand insurance that pays a benefit (in any state of the world) that is greater than W, because that kind of insurance would leave his wealth higher in states where that payment occurred, something he would not desire if he was risk-averse.

If W represents cash or liquid assets, this model is correct; the worst thing that could happen to someone is to lose all their wealth. If the person owns a physical asset, it is possible for a lawsuit associated with harm related to that asset to render a judgment greater than the value of the asset. The person will then declare bankruptcy, so in that sense cannot lose more than the wealth represented by what the physical asset is worth. But as many models of business demand for insurance note, there are non-monetary (or at least non-current-period) downsides to bankruptcy: time costs, shame costs, costs in terms of future career prospects. The person might then buy insurance to cover the cost of losses in excess of the value of the asset, if paying for that additional insurance coverage was preferred to experiencing these extra-wealth reductions in welfare. But note that the expected value of the insurance benefits will then exceed pW; observed payments will be greater with insurance than without, even though the insurer can perfectly distinguish the loss state of the world from the no loss state of the world.

The same sort of thing can happen with medical insurance. Suppose that there is small probability q that I may contract liver cancer, and I have wealth W. Without treatment, liver cancer is fatal within a short time period, but two treatments are possible: one treatment (say, chemotherapy) costs X, an amount which is less than W, and has a survival probability of p, and the other treatment (say, liver transplant) costs X' which is greater than both X and W, but has a higher survival probability p'. So my insurance choices are to buy a policy which pays X, costs qX', gives me expected survival probability of $(1-q)+pq$, and leaves my wealth at $W-qX$, while another policy costs qX, gives higher survival probability of $(1-q)+qp'$, and leaves me with smaller wealth $W-qX'$. I may prefer the second policy to the first, if the value of the increase in survival probability $q(p' -p)$ is greater than the value of the forgone consumption $q(X' -X)$. With no income effects, if insurance were unavailable, I would be forced to choose the treatment X, and have (p. 361) expected expense qX. With insurance possible, I would choose insurance that pays for the treatment X', and would have the larger expected expense qX'. There would be an observed positive effect on spending but (as in the wealth case in the previous paragraph) it would not come either from distorted incentives or from income effects, but from the fact that money wealth does not perfectly proxy all the things that go into a person's lifetime utility.

16.6 Moral Hazard and Optimal Health Insurance with Single-period Independent Demand

The classic model of moral hazard assumes a single period with households facing identical risks of different illness states. Relative to a benchmark “average health” state, some of the other possible health states represent the exogenous occurrence of illness for which there are costly but effective treatments available. Across a range of types of

medical care, the marginal health product of some or all of them is higher when a person is sicker; medical care (usually but not always) does you more good when you are sicker. More generally, the person is assumed to have a demand for medical care of the form:

$$D = D(P', Y - R, H'(H), Z)$$

Where P' is the marginal user price (the additional amount paid out of pocket) for a unit of a composite medical care, Y is disposable compensation (including the value of benefits), R is the health insurance premium, $H'(H)$ is the marginal health product (marginal benefit in terms of some indicator of health, such as healthy days), which depends on the person's health status H, and Z is a vector of other influences on the demand for medical care such as education, family size, etc. (I do not insert age in the demand function (in contrast to Grossman 1972, because I assume the effect of age is included in health status.)

The two variables in this demand expression that are influenced by health insurance (in the short run) are P' and R. The premium R is assumed to depend on the expected benefits from the coverage provided in the insurance policy, the insurer's "loading" for administrative expenses and profits, and any taxes, tax subsidies, or explicit subsidies for the insurance. In the US employment based health insurance system, R is often paid in two parts, as an explicit employee premium ("worker share"), and as a component of total compensation ("employer share").

The specification of P' can be complex. The simplest case is one in which insurance covers all components of the composite medical service, subject to a constant co-insurance rate c to be paid by the consumer as cost sharing. Then P' is just cP, where P is (p. 362) the gross unit price paid to the provider of the medical good or service. In more complex insurance designs some medical services may be excluded from coverage entirely, and there may be deductibles and upper limits on benefit payments for covered services. In what immediately follows we will ignore these complexities and assume that, in theory and in empirical estimation, the "quantity of insurance" will be represented solely by the co-insurance rate c.

The premium R will then depend on c, but it will also depend on the person's expected spending on medical care at any given level of P' (or c), and so on all of the variables that enter the medical care demand function. If the level of insurance coverage and the amount the person must pay for it is predetermined, one can then ask about the effect of various levels of c on the quantity of medical care demanded, and the great bulk of the literature treats the cost-sharing problem in this way. If the household has a choice of what insurance to buy, then both D and P' are obviously determined simultaneously (at least in an expectational sense), and so both D and P' are endogenous.

If income effects are zero, one can think of the effect of P' on D (given the assumption of a given gross price P that is large enough so that providers would make any quantities available) as equivalent to the quantities that would be read off a conventional (Marshallian) demand curve (that incorporates income effects) of price changes. To be

sure, the larger the quantity of care a given person demands, the higher the premium if premiums depend on expected benefits. But if the size of the pool of persons among whom the insurer is spreading the risk is reasonably large, the effect of this change on R (and hence on the person's demand for care) is negligible. However, the effect of changing P' on everyone's demand (or on the average insured person's demand) will be taken into account in determining the market premium for different levels of c. That is, the insurer will notice that, other things equal, the total amount of medical care demanded will be higher for people with lower values of C, and therefore price insurance coverage (and calculate the marginal premium for changing c) while taking this into account.

Given this type of effect on the use of medical care, how much and/or what type of insurance would the person choose? Let us take the simple case in which the insurance premium is actuarially fair. It is clear that potential insurance purchasers would prefer insurance that did not increase demand, or at least limited the increase in demand. One method of doing so is to increase the co-insurance proportion c. The other way of doing so is to have managed care that constrains the amount of care that will be covered to something less than what the person would demand with c equals zero, or some low level of c. In either case limiting moral hazard has a cost as noted, either in a greater risk of out of pocket cost or in a greater risk of not getting optimal care in the case of an illness of great but unobservable severity.

If only cost sharing can be used, it should be clear that, other things equal, the optimal limit on use will be greater the larger the responsiveness of use to insurance coverage. This is the “Zeckhauser proposition” (Zeckhauser 1970) that has been much discussed in the literature. In the case of linear demand curve, the key parameter turns out to be the slope of the demand curve; in the case of constant elasticity curves, it is the elasticity. In (p. 363) absolute value, the higher the slope or the elasticity, the higher the optimal level of co-insurance.

In the more general case, it will generally be preferable to use a combination of patient and provider “cost sharing” to limit moral hazard, rather than rely entirely on one or the other strategy (Ellis and McGuire 1993; Pauly and Ramsey 1999). The reason is that maximum quantity limits, even if imperfect, may be preferable to cost sharing in situations of high total spending, but cost sharing may work better to control moral hazard at lower levels of spending. The intuition is that limits constrain (high) spending close to the limit, while cost sharing offers incentives to control spending when spending is low. In addition, managed care which is limited to imposing upper limits on cost may be a good way to control moral hazard when demand responsiveness is high, but cost sharing may work better when demand responsiveness is low—thus reversing the Zeckhauser proposition about the relationship between optimal cost sharing and demand responsiveness.

16.7 Interrelated Demands

In reality there is no single or simple composite medical good or service. Instead, different types of care are used in different health states. Different services can be either close complements or close substitutes, which means that changes in the user price of one medical good can affect demand for other goods, conditional on their user prices. Sometimes these interrelationships are virtually contemporaneous, but sometimes the effect operates with a lag. In the lagged case, a medical good whose consumption now reduces my demand for some other medical goods in the future would generically be labeled “preventive care,” even though it may not be traditional preventive care. Other than issues of uncertainty and discounting, discussed below, such cost offsets capture what is sometimes called “dynamic moral hazard” (Zweifel and Manning 2000). The mechanism by which a good consumed in the present affects demand for a good in the future is usually through affecting the probability of future health states. This relationship is sometimes called “ex ante moral hazard” although it really can be viewed (for insurance purposes) as a special case of interrelated demands.

One difference between interrelated demands that occur within a short time period (like allergy shots and allergy attacks) versus those where the consequences are far in the future (like blood pressure medicine) is that in the latter case the size of the real discount rate is important. It is important for determining socially optimal coverage (at high real interest rates blood pressure control makes less sense) and in explaining patient behavior. As Newhouse 2006 has suggested, consumers may have irrationally high (hyperbolic) discount rates, higher than true real rates, making them less likely to see interaction effects as beneficial even if they have perfect information about the effectiveness of preventive care.

(p. 364) Once we consider the situation in which there are different kinds of medical goods and services, we need also consider the possibility that there are other goods, services, or activities not traditionally labeled “medical” which nevertheless affect health and therefore affect the demand for medical care. Exercising at gym or health club, for example, is a good which often has a positive price but which affects future health and therefore future demand for medical care.

The general proposition in this non-independent case is that insurance should cover a given type of medical care more generously, other things equal including the coverage of other medical services, the larger is the extent of substitution or cost offset. Conversely, the more generous the coverage of the “offset” services, the lower the optimal coverage of the offsetting service. Complementarity in demand will usually imply lower levels of coverage. Even activities not usually thought of as medical, like the cost of access to exercise facilities, might appropriately be covered, and coverage of a service with low or zero risk (like an annual preventive service that happens with certainty every year) might

be covered by insurance if it generates cost offsets. Generally it will be preferable to deal with cost offsets against a type of care initially generously covered by providing coverage to the offsetting service rather than by reducing coverage for the offset service, since the former provides more risk protection.

16.8 Insurance Coverage with Imperfect Supply-side Competition

The standard theory of insurance and moral hazard assumes that medical services are competitively supplied at prices that equal marginal cost, usually also assumed constant. If supply is not competitive and cannot construct average cost, then insurance effects on demand may translate into price changes. One simple case is that of a linear demand curve for care provided by a (constant marginal cost) monopolist; the lower the co-insurance rate, the higher will be the monopolist's profit maximizing price. In partial contrast, if the monopoly market demand curve were constant elasticity, then increases in the generosity of insurance coverage would leave the price unaffected since the monopolist's mark-up rule only contains the (assumed constant) values of marginal cost and demand elasticity; the only effect of insurance is to increase the quantity that the monopolist will sell at the given marked-up price.

That the monopoly quantity with moral hazard may be below the competitive quantity with moral hazard has raised the question of whether such over-pricing might offset the overconsumption that is associated with moral hazard. For a given exogenous value of insurance coverage there will be an offsetting effect of monopoly pricing. But Gaynor, Haas-Wilson, and Vogt (2000) have shown that if the insurance demander selects the level of coverage optimally, altering the product market from competitive to monopoly can never be welfare increasing. The intuition is that consumers will set cost sharing in (p. 365) the competitive case where the marginal welfare cost of changes in coverage just equals the marginal risk reduction benefit. Moving the market to monopoly at that co-insurance rate will lead to a sufficiently great increase in price that the consumer will choose to become exposed to more financial risk, and this increase in risk exposure more than offsets any savings from reduced consumption of medical care. That is, the optimal level of coverage at the higher monopoly price will leave cost sharing amount at a higher level than if the price were competitive, so that monopoly will mean both higher premiums and less risk protection.

16.9 Moral Hazard and Patterns of Voluntary Insurance Coverage: Theory and Practice

The forgoing discussion implies that consumers will demand the most generous insurance coverage against types of care or types of illnesses for which demand responsiveness is low, the probability of illness is low, and the cost of treatment is high. Conversely, coverage will be less generous if demand is price responsive and the illness is common and relatively inexpensive to treat.

If administrative expense were a fixed proportion of expected benefits, there would be positive amounts of insurance coverage for all uncertain medical services, but the generosity of coverage would vary as described above. If there is a fixed cost associated with initiating coverage for some new class of services, the presence of low risk and high price responsiveness may mean that no positive level of coverage is optimal.

This theory implies that, unless administrative costs inhibit and in the absence of adverse selection, there should be variation in the level of insurance coverage across services and illnesses, and there is some such variation, though it would be an exaggeration to say that there is as much variation in coverage as there is probable variation in price responsiveness and risk. One implication is that parity in coverage across different types of services or illnesses may be generally inefficient, and laws and regulations requiring parity may be welfare reducing.

In the theory just discussed, insurance coverage should not be uniform. Instead, cost sharing should vary with demand responsiveness and with administrative costs. In the voluntary market in the United States, and in other countries, coverage has historically varied across services. The general pattern is that inpatient care (both hospital and doctor) are almost completely covered. Outpatient care (in a doctor's office, hospital outpatient department, or clinic has modest cost sharing—for some reason, usually 20%). Prescription drugs, dental care, outpatient mental health care, and home health care are less well-covered, with average cost sharing in the 30 to 50 percent range. Nursing home (p. 366) care is barely covered by private insurance. Finally, very high (catastrophic) expenses usually have zero or minimal marginal cost sharing regardless of the type of service.

Part of the reason for this pattern in the United States is historical: Both hospitals and hospital-oriented physicians started what became the dominant insurance plans to cover their own services (Blue Cross and Blue Shield, respectively). Commercial companies typically offered “major medical” catastrophic coverage which was uniform across services with a high total cost (above a deductible). Commercial firms also pioneered coverage of prescription drugs, home health care, and long-term care.

As a broad generalization (except for nursing home care), it does appear that the pattern of coverage matches the theory; coverage varies directly with the size of the loss and inversely with demand responsiveness. However, until recently, insurances that used fine variations in the level of coverage to affect demand for specific services or in the case of specific illnesses had not occurred because of high administrative cost, but those costs

have been lowered in recent years by electronic processing. This is especially the case for coverage of prescription drugs which often had prohibitively high claims processing costs with paper-based methods, but which has taken the lead in electronic pharmacy claims-submission methods (Danzon and Pauly 2002).

The absence of voluntary coverage for nursing home or residential care is largely explained by crowd-out from the public Medicaid program (Brown and Finkelstein 2007, 2008). This program pays for nursing home care, but only after beneficiaries have spent their wealth and incomes. In other countries, long-term care is provided by a combination of social insurance, social services, and housing policy. There is no country with a large market in voluntary long-term care insurance. However, there are often continued disputes with the social insurance plan or trust about what will be paid, and consequent use of private funds if public support is thought to be inadequate. Requirements to use private wealth (including housing wealth, at some point) are common in other countries as well. In addition, the concentration of such expenses in a part of the life-cycle and health states where the marginal utility of other consumption is low may have depressed demand.

16.10 Empirical Evidence on the Effect of Insurance on the Demand for Medical Care

There has been extensive empirical investigation of the effect of changing cost sharing on consumer demand for medical care including, but not limited to, information from one of the most costly and most famous social experimental interventions, the Rand Health Insurance Experiment. It may not be an exaggeration to say that we know more about the elasticity of demand for medical services than for almost any other (p. 367) commodity in household budgets. With the amount of effort devoted to investigating this question, it is not surprising that we do have some definitive conclusions, but there are also some serious puzzles remaining, and more interesting new research continues to emerge.

The two main empirical findings from research to date are these: (1) the aggregate or average consumer demand curve, whether Marshallian (uncompensated) or Hicksian (compensated), slopes downward and to the right. (2) Demand curves are significantly price responsive at all consumer income levels. These conclusions are at variance with common perceptions of medical care demand by non-economists who traditionally have asserted that non-poor consumers only use medical care when they have to do so because they are sick or are ordered to do so by their physician, and that only lower income households would restrain their demand for needed care because of cost sharing. But there are also two serious questions which remain: (1) Does the effect of cost sharing just impact the consumer's decision to initiate care for an episode of illness, or does it also influence the rate at which care is used once a physician has been involved in the process; and (2) is the impact of changing cost sharing for an individual consumer the

same as the impact of a similar change for a large share of the population in a market? These two questions are related, because their answers depend on how physicians behave in response to patient cost sharing.

The question of the effect of insurance coverage of various types and amounts on the demand for medical services at various gross prices is obviously important to private or public insurance plans: it is crucial to understanding the benefits costs of changing coverage and, if combined with information on the value of coverage, is necessary (though not sufficient) for judging whether a change in coverage is efficiency improving and/or desirable to consumers or taxpayers. The most economically obvious way to set up the problem is to envision insurances with various levels of proportional co-insurance (including "no coverage" at 100%). The impact of a change in coverage which changes co-insurance would be expected to be very similar to the impact of a single change in gross price. That is, we might assume that the effect of change in the price of a doctor visit from $100 to $50 on the quantity of care demanded with no insurance is the same as the effect of going from no coverage to 50 percent co-insurance with the gross price of a visit remaining at $100 (except for the possible income effect of paying the premium in the latter case). Moral hazard in this sense is nothing more than movement along an ordinary demand curve.

There has been extensive investigation of the impact of changing insurance coverage on quantities of specific medical services and on total medical care spending; there has been much less investigation of any effects on the quality of care (except as captured by a movement to higher priced and presumably higher quality services). Finally, there has been some investigation of the effect of changes in insurance coverage on health outcomes.

It will be useful to use as a benchmark the most discussed and most expensive empirical estimates of medical care demand elasticities, those from the Rand health insurance experiment (Newhouse and The Insurance Experiment Group 1993). That experiment, **(p. 368)** conducted in the United States in the early 1980s, was a social experiment that randomly assigned samples of the US population under the age of 65, non-institutionalized, and with incomes below an upper income cut-off to a variety of different health insurance plans, ranging from "free care" (coverage without cost sharing for medical, dental, and eye care) to plans with varying levels of co-insurance and deductibles. The experiment did not include a subpopulation with no insurance. While I will provide some more details below, the key parameter values generated by the experiment were estimates of demand elasticity that were statistically significantly different from zero, and in the range of -0.1 to -0.2.

While the experiment has never been repeated, more recent estimates of demand elasticity generated by natural variation in coverage that have tried to deal with the problem of endogeneity, have confirmed the Rand result of a negatively sloped demand curve for almost all populations in almost all settings. Generalizations are difficult here,

but it is my impression that these more recent studies have rarely found point estimates lower than those from the experiment. The best current consensus estimate of the elasticity is at or above the Rand estimate.

Before discussing the actual estimates of demand elasticity, I consider the key issue of endogeneity of coverage that has figured so strongly in empirical modeling strategies. To begin with, one of the main rationales for the Rand experiment was that it was a design which should automatically avoid problems of endogeneity by the use of the clinical trial model with random assignment. The rationale for this was that earlier observational studies may have yielded biased estimates. While bias is a potential problem any time there is consumer choice, it is helpful to think about what the nature and direction of the bias might be in order to interpret estimates where there is potential bias and to judge whether it is worthwhile to go to greater effort to avoid bias. Early estimates of the relationship between insurance coverage and medical care use or spending generally used cross sectional data where observations were households or geographic areas. Particularly in the former case, if adverse selection was likely, the observed relationship between coverage and use or spending would be biased upwards in magnitude as an estimate of moral hazard alone. High risks would have more insurance (away from zero), and the high spending by the well-insured could be caused by their higher risk as well as by their generous coverage. (Indeed, a recent spate of econometric studies intended to estimate adverse selection have had to deal with moral hazard as a possible contaminating influence (Cardon and Hendel 2001; Town 2008).) The implication here is that the bias will be greater the greater is adverse selection. Adverse selection will bias downward (toward zero) estimates of the relationship between coverage and health outcomes, because the higher risks who selected coverage would probably have had worse health outcomes than the average risks (Pauly 2005).

In contrast, if insurers are able to identify high risks (thus avoiding adverse selection) and charge them higher premiums, this might cause high risks (at least, low-income high risks) to forgo coverage or to buy less generous coverage. Or, as some recent research indicates, people with higher levels of financial risk aversion have lower levels of current health risk because they have avoided risk taking behavior and consumption **(p. 369)** items that threaten health. In these cases, the observed relationship between coverage and spending would be a downward biased estimate of moral hazard and the observed relationship between coverage and health outcomes would be biased away from zero. As will be discussed in more detail below, there is evidence of both kinds of selection in voluntary insurance markets.

The early studies that did not control for endogeneity did indeed sometimes get empirical estimates of demand elasticity which were sometime much larger than the Rand range, but they also sometimes got estimates below -0.1 (especially when they looked at household data when adverse selection was most likely). If one takes the Rand estimates as an unbiased benchmark, it is hard to generalize about the direction or magnitude of bias in these earlier studies.

I now briefly review the results of the Rand experiment and subsequent discussions of interpretation of those results. The experiment assigned people to insurance plans with varying levels of co-insurance (0, 25, 50, and 95%) with stop-losses at a predetermined percentage of household income. As noted, use and expenditures on medical care were significantly higher under the free care (0% co-insurance) plan than under the cost sharing plans; the increment in average expense from replacing cost sharing with free care was as high as 46 percent. All types of medical care spending were higher in the free care setting: inpatient care, outpatient care, preventive care, prescription drugs, and dental care. The elasticity was less than average in magnitude for inpatient care, especially for children, and was higher than average for dental care, preventive care, and outpatient mental health care. (Since the plans generally imposed uniform co-insurance, Rand results cannot provide estimates of the effect of changing co-insurance for one type of care while holding the others constant—except for an outpatient deductible plan which was found to discourage both outpatient and inpatient care, implying that they were gross complements.) The response to cost sharing did not differ by household income or other household characteristics, although there was a small income effect on the use of care. It appears that the primary impact of cost sharing was on the rate of initiation of episodes of care; expense per episode did not differ across plans. That is, once care is initiated and is being guided by a physician, variations in individual patient cost sharing appear to have little impact: doctors treat everyone the same.

The experiment also looked at a set of indicators of health outcomes that were supposed to be sensitive in the short run to the use of care. The only indicators that were affected by cost sharing for all population groups were measures of oral health and vision correction. For the low-income group originally at high risk, free care was associated with better blood pressure control, but there were no significant effects for middle income households or for low-income households at average or better risk. The pattern of minimal effects on health was not always consistent with professional judgments about the effectiveness of care. People with cost sharing did use emergency room care less frequently for causes that were not true emergencies (compared to those that were true emergencies), but there was no differential effect of cost sharing on other ambulatory or inpatient care labeled as medically necessary or appropriate; cost sharing discouraged (p. 370) care to the same extent whether it was thought to be effective for health outcomes or not. There was no investigation of the use of care characterized as having positive but low effectiveness (in absolute terms or relative to cost). Since there is no theoretical expectation in demand theory that people will use care that is ineffective or harmful, even if it is free, the mixed results for health effects remain puzzling. One would have to invoke supplier-induced demand that got stronger as cost sharing fell to explain these results.

Empirical work since the experiment has been of four types: estimates of the impact of no insurance relative to typical insurance coverage; attempts to replicate the experiment using other methods for assuring exogeneity of coverage; explorations of cost sharing in

plans that also contain managed care features, and research on settings in which coverage was varied for a large fraction of the consumers in a market, rather than just for a handful as in the experiment.

A recent survey article by Buchmueller et al. (2005) looked at the US literature regarding the effect of having some insurance versus having no insurance on use of outpatient and inpatient care. If we assume that insurance on average covered 80 percent of outpatient care cost and 90 percent of inpatient care cost, the implied elasticities are −0.4 to −0.8 for outpatient care, and −0.25 to −0.5 for inpatient care. The general pattern is thus one of elasticities that are definitely negative and significant with numerical values somewhat higher than those for the experiment.

An example of the second kind of study is an analysis by Matthew Eichner 1998 that used as the instrument for lowered cost sharing the exogenous impact of care for a family member's accident in insurance plans with family deductibles. The goal was to examine how care was used after the deductible was covered (implying zero or low cost sharing for any additional care) compared to households with positive expected cost sharing because they were still liable for some of the deductible. This work found a statistically significant effect of free care on use and spending; the point estimate of the elasticity was in the range of −0.7.

A third set of studies look at the effects of cost sharing that differ from the uniform co-insurance, fee-for-service model of the Rand experiment. One study by Hillman et al. 1999 looked at the effect of different levels of drug co-payment per prescription in two kinds of managed care settings: one was a fee-for-service IPA model where physicians were at risk for the cost of their own services but not for the cost of drugs they prescribed, and the other a network model where doctors were at risk (collectively) for their prescription drug costs. Drug co-payments varied as cost sharing for other services held constant. One might have expected some supply-side restraint on prescribing compared to non-managed care fee-for-service even in the first case, since physician visits and drug prescriptions are complements, but one would have expected more restraint in the second case. This was indeed the outcome: the frequency of cost of prescriptions was lower at higher co-payments than at lower co-payments, but the patient responsiveness to cost sharing was much greater in the first setting than in the second. The point estimate of elasticity in the fee-for-service setting was a little more than −0.2, but was much smaller in the second setting.

(p. 371) Other studies of drug cost sharing have looked at the effect of so called “triple tier” design where cost sharing is highest for non-preferred brand name drugs. The almost universal finding is that such cost sharing shifted prescribing and use away from the higher co-pay drugs. Its impact on overall drug spending depends on the comparator: if the comparator is a plan with a closed formulary (so no coverage at all for non-preferred drugs) it led to higher spending (Rector et al. 2003); but if the comparator was

lower and uniform across the board set of co-payments there was a reduction in total use as well, not just a shift in patterns of use.

The behavior observed in these first three kinds of studies obviously depends on more than just consumer choice, since the consumer needs a physician prescription or order to obtain a prescription drug or many other non-emergency services. The broader issue of how physician behavior seems to change in response to changes or variation in patient cost sharing is just now beginning to be examined. Some older hypotheses about physician behavior suggested that physicians try to achieve a target income, or at least try to offset the impact of external changes which reduce their income (Evans 1974). Simple versions of this hypothesis are obviously inconsistent with the finding that cost sharing reduces the use of physician services or services from which physicians benefit; if the target income hypothesis were literally true, doctors would increase patient demand enough by inducement to off set the effects of cost sharing. But there is some evidence for physician inducement especially in response to exogenous reductions in the gross price for their services (Yip 1998). Probably both inducement and cost sharing affect the quantity that ends up being provided. But how the exchange of information, physician orders, and patient psychology interact to produce this blended result is unknown.

Another insight that incorporates physician behavior notes that physicians may well respond differently to changes in their patients' cost sharing when the change occurs for a large number of patients (as often happens in real markets) rather than for just a tiny fraction of patients (as in the Rand experiment and the cross sectional studies using individual data). A plausible assumption is that physicians would prefer to treat all of their patients with a given illness of a given severity approximately the same; this is less costly (in monetary and effort terms) than paying attention to each patient's insurance coverage (Glied and Zivin 2002). There are costs to “pattern of practice” differentiation in a physician practice as in other markets. And it may be ethically uncomfortable to treat differently based on insurance rather than on illness state only. Both of these reasons have been offered as explanations for the Rand finding that cost sharing did not affect the cost per treated episode (although a censoring story, in which people with high sharing only seek treatment for severe illness, could also explain that result). There is evidence that the care a given patient gets may depend on the insurance coverage of others in the local market as well as on that person's insurance status (Pauly 1979; Pagán and Pauly 2006, 2005). The strongest evidence that demand response is greater for a large-scale change is offered by a recent study by Finkelstein 2007 that looked at the effect of the introduction of Medicare in the United States (which dramatically lowered average cost sharing); the estimated response in terms of both use and spending was much greater than would have been implied by the Rand experiment elasticities. The response (p. 372) was especially large for newer technologies which would have needed a large total market to be profitable.

16.11 Insurance Deductibles and the Use of Medical Care

Public and private health insurance frequently contains deductibles—money amounts per time period (or, occasionally, per illness) that must be paid out of pocket before any insurance benefits are paid. In the pure theory of insurance, Arrow 1963 showed that, with proportional administrative loading, optimal coverage is full coverage above a deductible. The intuition behind this conclusion is that the marginal risk premium for losses relative to the mean becomes vanishingly small as the size of the loss shrinks, and therefore becomes less than the marginal loading, which remains positive. This rationale for a deductible is only strengthened if the administrative cost structure also has a positive marginal cost per claim; insurance will not be worthwhile for a set of small claims amounting to a small total amount.

When the consumer has price-sensitive demand for care, the influence of deductibles on spending is complex because a deductible in effect faces the consumer with a two part block tariff: full price up to a certain level of total spending, and then low or zero marginal price. Since the marginal price is different depending on whether the deductible is covered or not, the consumer has to consider the distribution of expected expenses (See Zweifel and Manning 2000 for a discussion of such models). While the actual analytics of demand responsiveness are complicated by a deductible—since the relevant empirical price depends on the (expected) distribution of expenses over the different ranges of marginal prices—the main intuitive finding is obvious: the lower the deductible the higher the demand for care, other things equal.

There has been some policy interest in the effects of deductibles on spending because of the controversy about high deductible health plans/health savings accounts arrangements in the United States and their supposedly large effects on demand. One issue is the impact of a deductible when the cost of a treated illness is sure to exceed the deductible. Some critics of high deductible plans point out that, regardless of the size of the deductible, the *ex post* marginal price will be zero, and conclude that there should be little effect on demand from high deductibles and that they would expose people to higher financial risk. However, this argument only follows if the realized zero marginal price is the only one facing all potential buyers of high cost treatment. Somewhat surprisingly, the Rand experiment showed that this conjecture is empirically incorrect and that cost sharing does matter for the use of high-cost treatment. The reason is probably because cost sharing affects the consumer/patient's decision to initiate treatment that will turn out to be high cost. Think of an inpatient hospital admission for some illness; any inpatient care in the United States is virtually certain to have a high cost that exceeds the maximum (p. 373) permitted deductible in tax subsidized high deductible health plans. But the higher the deductible, the higher the out of pocket cost of initiating that episode of care. An increase in the deductible does not have to discourage many admissions to save a lot (although the Rand experiment does say that low deductibles have the

strongest marginal effect on total medical spending). A more correct analysis would exempt hospital admissions known to be virtually non-discretionary (like care following a heart attack or serious burn accident) from high deductibles.

The other controversial issue concerns the impact of high deductibles on the use of prevention and early treatment. The conventional wisdom is that high deductibles or exclusions will thereby discourage their use. Some high deductible plans, in response to this concern, exempt a small number of highly effective preventive services from the deductible. The higher user price from a high deductible would, in isolation, discourage the use of preventive or early care that might have reduced the cost of some illness. But a key question is the extent of patient cost sharing should the illness occur. If the deductible is high enough that a large portion of the cost of the potential illness is also under the deductible, there should be an offsetting and potentially strong incentive to pay out of pocket for prevention or early care, since it avoids yet higher out of pocket payments later on. (If the cost of an episode of flu will still be under the deductible, it will pay for me to get a flu shot even if it is also not covered.) Conversely, if most of the cost of the illness is above the deductible, there will be a weaker incentive to use preventive care. So one cannot generalize a priori.

16.12 Moral Hazard and Value-based Cost Sharing: Theory and Evidence

There are many interesting threads introduced by consideration of physician—patient interactions. One particular issue concerns the level of patient knowledge of marginal benefit from care. A conventional demand model (as applied to medical care or anything else) assumes that the buyer makes some estimate of the (marginal) benefit from various quantities of various goods in deciding how much to demand. In principle one could specify a level of information that is less than complete as an additional demand influence; in simple modeling consumers are usually assumed to be as well-informed as is possible, and to behave according to their well-informed demand curves.

If patients do not have correct perceptions of the marginal benefit from consumption of medical care, they may make incorrect decisions. There may therefore be scope for altering cost sharing away from the level that would have been ideal with correct information and toward the level that would push people to the right choice. An alternative is to provide them with correct information. There are three reasons why a consumer/ patient may not have correct perception of the marginal benefit from some care: (1) the patient may be deciding whether or not to initiate an episode of care based on symptoms (p. 374) and knowledge that do not allow the patient to know whether treatment of the symptoms or illness is beneficial, or beneficial enough relative to cost; (2) having initiated an episode of treatment by contacting a physician, the physician may not have been able to convey or convince the patients about the marginal benefit from

care; (3) if use of care has interaction effects on demands for other kinds of care, insurance coverage of the costs of those other kinds of care may disguise the expected total correct marginal cost of close substitutes or complements.

Let us consider the third case in the context of preventive care which causes cost offsets (reductions) in the use of other medical care. Assume that the patient is correctly informed about the health benefits (in terms of reduced future probability of illness). If there was no insurance coverage for either “prevention” or “treatment” (say, because the cost of either or both fell below the deductible on a policy), then the patient would face optimal incentives: pay for the preventive care if the combination of the value of improved future health and future treatment cost savings is greater than the cost of the preventive service. In contrast, if insurance covered all of the cost of treatment but none of the cost of prevention, in deciding on consumption of preventive care the patient would ignore the cost savings. One way to improve incentives would be to offer treatment coverage at reduced premiums for those who had already bought preventive care, but another way would be to have insurance cover some of the cost of preventive care. The ideal incentive would reduce the price of the preventive treatment by the expected value of the cost offset.

So far, the issue of the demand for prevention has not arisen. If the preventive service is binary—you either get a flu shot or you do not—the shape of the demand curve for prevention from a given population of people insured for treatment depends on the distribution of reservation prices for prevention. This in turn will depend on any non-monetary cost for the preventive service—values consumers place on pain, time, or inconvenience, and the value they place on avoiding the adverse health outcome, given the treatment they would expect to consume. If reservation prices are all above the value of the improved health outcome, there is no gain from lowering the user price of preventive care, since all would use it anyway. If some reservation prices are below the full cost of the preventive service, there will be a larger gain at any level of coverage the more price responsive is demand; this is the benign moral hazard discussed by Pauly and Held 1990. Gain is maximized at any level of demand elasticity by setting the incentive at the optimal level, but the amount of gain (and therefore the offset against any transactions cost from a subsidy) is greater the higher the price responsiveness.

Now let us consider the case in which there is no cost offset but patients have not been correctly informed about marginal benefit (or the discount rate for future benefits). With correct information, the more elastic the schedule of marginal benefit, the higher is the optimal level of co-insurance. If marginal benefit is uniformly under-estimated, Pauly and Blavin 2008 show that co-insurance rates should be lower than under correct estimation, but should still vary inversely with demand elasticity or price responsiveness. In the case of over-estimation of marginal benefit, co-insurance should be higher than in the correct information case but also vary inversely with responsiveness. One additional **(p. 375)** finding of interest: if it is costly to correct under-estimation of marginal benefit, it may be preferable to use lower co-insurance (which also provides better risk protection) to move

use closer to the optimal point than to provide information. The reason is that reducing co-insurance increases risk protection, and so has a negative cost, whereas information provision has a positive cost.

16.13 Moral Hazard and New Technology

Medical care spending rises in developed countries largely because of changes in technology which improve the quality of care but at a higher net cost. Higher levels of cost sharing reduce cost and use when they are implemented, but have no theoretically predictable effect of the rate of growth of spending. This is because, without more specific assumptions, we should assume that they reduce the rate of use of any new technology by the same proportional amount as they reduce the base of use of older technology; with equiproportional reductions in base and increment, the rate of growth remains the same.

Empirical evidence on techniques to reduce moral hazard in new technology has not yet produced definitive results. Aggressive managed care, such as that embodied in staff or group model HMOs, seems to reduce use and spending below fee-for-service counterparts with the same (very small or zero) cost sharing, but not compared to coverage with typical cost sharing, and it does not seem to reduce the rate of growth in spending across the board. There is some evidence from US aggregate data that lower overall proportions of out of pocket payment are associated with higher rates of growth of spending (Peden and Freeland 1998). A complete model which determines the rate of investment in technological change, the rate and form of introduction of new technology, and the ideal level or pattern of insurance cost sharing has yet to be determined. Insurers could potentially use cost sharing and coverage to select different rates of growth in spending and technology, especially if insurers retain the power to refuse to cover (set co-insurance below 100%) for new technologies whose adoption under lower levels of co-insurance might make insured populations worse off (Goddeeris 1984a and 1984b). Definitive results on the actual or ideal relationship between insurance coverage and spending growth have yet to be established.

16.14 Beneficent Moral Hazard in Social Insurance

While the bulk of the normative economic theory dealing with the impact of insurance on the use of care views such an impact as having a negative effect both on welfare and the demand for insurance, health policy discussion by policymakers and health (p. 376) advocates frequently views patient cost sharing with considerable apprehension, and not just because it is a necessary evil to control spending. Instead, public policy in all

countries views both health insurance and the increase in use (or “improvement in access”) associated with such insurance as positively desirable. Obviously there is not a large volume of economic research consistent with this view, but are there economic interpretations and theories (beyond the cost offset or patient adherence considerations) that would conclude that cost sharing lower than that suggested by insurance theory is socially desirable?

One approach that leads to such a conclusion begins with the observation or assumption that there may be external benefits at the margin from the use of medical services and goods in excess of that which uninsured consumers would choose. Externalities can come from two main sources: contagious disease (Phelps 2003) and altruistic concern for others' health or care use (Pauly 1971). Given the relatively small share of spending attributable to preventable or treatable contagious disease in developed countries, the second rationale is of potentially greater quantitative significance. The fundamental model can be described in words: add the demand or marginal valuation of care by others in the community to that of each individual, equate the summed marginal evaluation to marginal cost to define the optimal quantity, determine the user price (below market price) at which demand equals the optimal quantity, and use subsidies, mandates, or regulations to make sure that insurance lowers user price to at least this level. For the well-off, the insurance and care they would demand without subsidies is likely to display little or no external benefit at the margin-they use “enough” care-but especially for low-income households the level of coverage may well be determined at the margin by this care-specific altruistic externality (Folland, Goodman, and Stano 2009).

Despite the obvious importance of social insurance in determining the insurance coverage people have and the obvious importance of this theory in specifying the optimal form of social insurance, there is relatively little research on it. The rationale for social insurance is sometimes specified instead as dealing with market failure, especially the possibility of adverse selection (Zweifel, Breyer, and Kiffmann 2009), or as a version of optimal taxation theory (Petretto 1999); in these cases the more traditional moral hazard theory would apply. In a study to present the case for universal health insurance, the Institute of Medicine (2003) measured benefit by human capital (discounted future earnings attributed to health improvements associated with insurance), rather than appealing to any social benefits.

There has been some work relating choice of generosity of coverage to taxpayer demand based on altruism (Grannemann 1980; Holahan and Chang 1989) but no direct test of the connection between coverage, private demand, and public demand. One implication of this theory is that a necessary condition for optimal coverage is that the medical care use induced by lower cost sharing is still cost effective at the margin, where the value of health is the sum of the patients' and society's willingness to pay. There have been some attempts to monetize the societal willingness to pay but the distinction between private benefits (I want to pay taxes for Medicare because I will value my improved health outcomes when I am on Medicare) and social benefits (I want to pay (p. 377) taxes for

Medicare because I will value improved health outcomes for old people other than myself) is usually not made.

One possible explanation for the lack of attention to the design of insurance in this context is that the most prominent target population, poor households, would demand socially suboptimal levels of care even when there is no cost sharing and coverage is complete. In US Medicaid, for example, cost sharing is usually zero or nominal at best. Any limitation on volume of care is produced by changing the level of provider fees or other influences on access, rather than by manipulating buyer demand. When income rises high enough for cost sharing to be relevant, perhaps social concern at the margin (except for very high risk people) is close to zero. Or more generally, social insurance systems may prefer to use supply side incentives to affect quantities rather than insurance coverage per se.

16.15 Conclusion

Insurance coverage does affect the use and cost of medical care, and so potentially can play a role in assuring that spending comes closer to the (second best) optimum. (The first best optimum requires so-far infeasible indemnity insurance.) But the information for making judgments about the value of care whose use is discouraged by cost sharing has so far not proven universally definitive and not even persuasive where it is reasonably definitive. Disentangling the intricate web of insurance coverage for medical services with interrelated demands and with demand subject to different levels and types of misinformation is a daunting challenge, but one worth accepting.

References

AUSTVOLL-DAHLGREN, A., AASERUD, M., VIST, G., RAMSAY, C., OXMAN, A., STURM, H., KÖSTERS, J., and V**ERNBY**, A. (2008). Pharmaceutical policies: effects of cap and co-payment on rational drug use. *Cochrane Database of Systematic Reviews*, 1: Art. No. CD007017, DOI:10.1002/14651850.CD 007017.

A**RROW**, K. J. (1963). Uncertainty and the welfare economics of medical care. *American Economic Review*, 53(5), 941–73.

B**ROWN**, J. R., and F**INKELSTEIN**, A. (2007). Why is the market for long-term care insurance so small? *Journal of Public Economics*, 91(10), 1967–91.

—— (2008). The interaction of public and private insurance: Medicaid and the long-term care insurance market. *American Economic Review*, 98(3), 1083–102.

BUCHMUELLER, T. C., GRUMBACH, K., KRONICK, R., and K**AHN**, J. G. (2005). Book review: the effect of health insurance on medical care utilization and implications for insurance expansion: A review of the literature. *Medical Care Research and Review*, 62(1), 3–30.

C**ARDON**, J. H., and H**ENDEL**, I. (2001). Asymmetric Information in health Insurance: evidence from the National Medical Expenditure Survey. *RAND Journal of Economics*, 32(3), 408–27.

(p. 378) **CHERNEW, M. E., ROSEN, A. B.**, and F**ENDRICK**, M. A. (2007). Value-based insurance design. *Health Affairs*, 26(2), w195–w203.

D**ANZON**, P., and P**AULY**, M. V. (2002). Health insurance and the growth in pharmaceutical expenditures. *The Journal of Law and Economics*, XLV (2, Part 2), 587–613.

DE MEZA, D. (1983). Health insurance and the demand for medical care. *Journal of Health Economics*, 2(1), 47–54.

E**ICHNER**, M. J. (1998). The demand for medical care: what people pay does matter. *American Economic Review*, 88(2), 117–21.

E**LLIS**, R. P., and M**CGUIRE**, T. G. (1993). Supply-side and demand-side cost sharing in health care. *Journal of Economic Perspectives*, 7(4), 135–51.

E**VANS**, R. G. (1974). Supplier-induced demand: some empirical evidence and implications. In M. Perlman, ed., *The Economics of Health and Medical Care*(pp. 162–73). Edinburgh: Macmillan.

F**INKELSTEIN**, A. (2007). The aggregate effects of health insurance: evidence from the introduction of Medicare. *Quarterly Journal of Economics*, 122(1), 1–37.

FOLLAND, S., GOODMAN, A., and S**TANO**, M. (2009). *Economics of health and health care*, 6th edition (pp. 392–94). Upper Saddle River, NJ: Prentice Hall.

GAYNOR, M., HAAS-WILSON, D., and V**OGT**, W. B. (2000). Are invisible hands good hands? Moral hazard, competition, and the second-best in health care markets. *Journal of Political Economy*, 108(5), 992–1005.

—— L**I**, J., and V**OGT**, W. (2006). Is drug coverage a free lunch? Cross-price elasticities and the design of prescription drug benefits. NBER Working Paper No. 12758.

G**LIED**, S., and Z**IVIN**, J. G. (2002). How do doctors behave when some (but not all) of their patients are in managed care? *Journal of Health Economics*, 21(2), 337–53.

G**ODDEERIS**, J. H. (1984a). Insurance and incentives for innovation in medical care. *Southern Economic Journal*, 51(2), 530–39.

—— (1984b). Medical insurance, technical change and welfare. *Economic Inquiry*, 22(1), 56–67.

G**RANNEMANN**, T. (1980). Reforming national health: Programs for the poor. In Mark Pauly, ed., *National Health Insurance: What Now? What Later? What Never?*Washington, DC: American Enterprise Institute.

G**ROSSMAN**, M. (1972). On the concept of health capital and the demand for health. *Journal of Political Economy*, 80(2), 223–55.

HILLMAN, A. L., ESCARCE, J. J., RIPLEY, K., GAYNOR, M., CLOUSE, J., and R**OSS**, R. (1999). Financial incentives and drug spending in managed care. *Health Affairs*, 18(2), 189–200.

H**OLAHAN**, J., and C**HANG**, D. (1989). *Medicaid Spending in the 1980s*. Washington, DC: Urban Institute Press.

INSTITUTE OF MEDICINE (2003). *Hidden costs, value lost: uninsurance in America*. Washington, DC: Committee on the Consequences of Uninsurance, Institute of Medicine.

MA, C.-T., and M**CGUIRE**, T. (1997). Optimal health insurance and provider payment. *The American Economic Review*, 87(4), 685–704.

N**EWHOUSE**, J. P. (2006). Reconsidering the moral hazard-risk avoidance trade-off. *Journal of Health Economics*, 25(5), 1005–14.

—— and **THE INSURANCE EXPERIMENT GROUP** (1993). *Free for all? Lessons for the Health Insurance Experiment*. Cambridge, MA: Harvard University Press.

N**YMAN**, J. (1999). The economics of moral hazard revisited. *Journal of Health Economics*, 18(6), 811–24.

(p. 379) **PAGÁN, J. A.**, and P**AULY**, M. V. (2005). Access to conventional medical care and the use of complementary and alternative medicine. *Health Affairs*, 24(1), 255–62.

—— (2006). Community-level uninsurance and the unmet medical needs of insured and uninsured adults. *Health Services Research*, 41 (3, Pt 1), 788–803.

P**AULY**, M. V. (1968). The economics of moral hazard. *American Economic Review*, 58(3), 531–37.

—— (1971). *Medical Care at Public Expense: A Study in Applied Welfare Economics*. New York: Praeger Publishers, Inc.

—— (1979). *Doctors and Their Workshops*. Chicago: University of Chicago Press.

—— (2005). Effects of health insurance on use of care and outcomes for young women. *The American Economic Review*, 95(2), 219–23.

—— (2008). Adverse selection and moral hazard: implications for insurance markets. In F. A. Sloan and H. Kasper, eds., *Incentives and Choice in Health and Health Care*(pp. 103–29). Cambridge, MA: MIT Press.

—— and B**LAVIN**, F. E. (2008). Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics*, 27(6), 1407–17.

—— and H**ELD**, P. J. (1990). Benign moral hazard and the cost effectiveness of insurance coverage. *Journal of Health Economics*, 9(4), 447–61.

—— and R**AMSEY**, S. D. (1999). Would you like suspenders to go with that belt? An analysis of optimal combinations of cost sharing and managed care. *Journal of Health Economics*, 18(4), 443–58.

P**EDEN**, E. A., and F**REELAND**, M. S. (1998). Insurance effects on US medical spending (1960–1993). *Health Economics*, 7(8), 671–87.

P**ETRETTO**, A. (1999). Optimal social health insurance with supplementary private insurance. *Journal of Health Economics*, 18(6), 727–45.

P**HELPS**, C. E. (2003). *Health Economics*, 3rd edition. New York: Pearson Addison-Wesley.

RECTOR, T. S., FINCH, M. D., DANZON, P. M., PAULY, M. V., and **MANDA, B.S.** (2003). Effect of tiered prescription co-payments on the use of preferred brand medications. *Medical Care*, 41(3), 398–406.

THOMSON, S., MOSSIALOS, E., and J**EMIAI**, N. (2003). *Cost Sharing for Health Services in the European Union*. London: London School of Economics, LSE Health and Social Care.

T**OWN**, R. J. (2008). Adverse selection, welfare and the optimal pricing of employer-sponsored health plans. Paper presented at Leonard Davis Institute Research Seminar, December 3, 2008, Wharton School, University of Pennsylvania, Philadelphia, PA.

W**AGSTAFF**, A. (2008). Measuring financial protection in health. World Bank Policy Research Working Paper No. 4554.

Y**IP**, W. C. (1998). Physician response to Medicare fee reductions: changes in the volume of Coronary Artery Bypass Graft (CABG) surgeries in the Medicare and private sectors. *Journal of Health Economics*, 17(6), 675–99.

Z**ECKHAUSER**, R. (1970). Medical insurance: A case study of the trade-off between risk spreading and appropriate incentives. *Journal of Economic Theory*, 2(1), 10–26.

Z**WEIFEL**, P. J., and M**ANNING**, W. G. (2000). Moral hazard and consumer incentives in health care. In A. J. Culyer and J. P. Newhouse, eds., *Handbook of Health Economics*, Volume 1A (pp. 409–59). Amsterdam: Elsevier.

—— B**REYER**, F., and K**IFMANN**, M. (2009). *Health Economics*, 2nd Edition. New York: Springer-Verlag New York, LLC.

Notes:

(1) Pauly 1968.

Mark V. Pauly

Mark V. Pauly is Bendheim Professor and Professor of Health Care Management at the Wharton School of the University of Pennsylvania. He previously held a faculty appointment at Northwestern University. Pauly currently serves as the co-editor-in-chief for the International Journal of Health Care Finance and Economics and as an advisory editor for the Journal of Risk and Insurance. Professor Pauly has been a member of several scientific panels, including the Medicare Technical Advisory Panel, the National Advisory Council for the Agency for Healthcare Research and Quality, and the Committee on Evaluation of Vaccine Purchase Financing in the United States at the National Academy of Sciences. Pauly has also served as a consultant for pharmaceutical companies, health care organizations, and public policy think tanks. Professor Pauly's current research includes projects that explore health care reform, conceptual foundations for cost-benefit analysis of drugs, and incentives in managed care. He received an AB from Xavier University, an MA from the University of Delaware, and a PhD in economics from the University of Virginia.